2020

Essays On Algorithms, Markets, And Society

Hadi Elzayn
University of Pennsylvania

Essays On Algorithms, Markets, And Society

Abstract

This thesis examines algorithmic markets - market mechanisms with algorithms as a core component in their functioning -- and markets with algorithms (that is, canonical markets for algorithmic or data-based goods and services). Our primary focus is on the analysis of these mechanisms and markets in terms of societal concerns such as fairness, privacy, and efficiency (including welfare and revenue). For algorithmic markets, we consider automated ad auctions and call auctions; for markets with algorithms, we examine both an abstract, general data-driven market from a theoretical perspective, and a specific, important data-driven market: the U.S. mortgage market. We apply a variety of theoretical tools from various subfields of Computer Science and Economics, including worst-case asymptotic runtime and sample complexity theory, the Probably Approximately Correct (PAC) Learning framework, no-regret learning algorithms, equilibrium analysis, smoothness, differential privacy, and quantitative fairness. In addition to theoretical analysis, we implement various algorithms and mechanisms and perform empirical analysis on real data in relation to the mortgage market. Our results include: new worst-case welfare guarantees, novel equilibrium characterizations, and experimental evaluation of various auction formats in the Ad Types setting; construction and analysis of a differentially private call auction mechanism with good performance and incentive properties; a theoretical analysis elucidating the economic forces that encourage error inequality in data-driven markets; and practical application of quantitative fairness measures to a uniquely rich and exhaustive dataset covering the mortgage market in the United States.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Applied Mathematics

First Advisor

Michael Kearns

Keywords

Algorithms, Auctions, Fairness, Learning, Markets, Privacy

Subject Categories

Computer Sciences | Economics

ESSAYS ON ALGORITHMS, MARKETS, AND SOCIETY

Hadi Stephen Elzayn

A DISSERTATION

in

Applied Mathematics and Computational Science

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2020

Supervisor of Dissertation

DocuSigned by:
Michael Kearns
5459BF180BE8492...

Michael Kearns, Professor of Computer and Information Science and National Center Chair

Graduate Group Chairperson

Robin Pemantle, Merriam Term Professor of Mathematics

Dissertation Committee:
Michael Kearns, Professor of Computer and Information Science and National Center Chair
Aaron Roth, Professor of Computer and Information Science
Sampath Kannan, Henry Salvatori Professor of Computer and Information Science

ACKNOWLEDGMENT

My time in graduate school has been thrilling, humbling, and certainly challenging. I have been privileged to work with brilliant colleagues and mentors on a wide breadth of fascinating topics, and there is no question that I am greatly indebted to many that have helped and been there for me along the way.

I have been extremely fortunate to be advised by Michael Kearns; his ability to conceptualize, explain, and make concrete even the most technical of problems has been inspirational, and his insight and guidance have helped me not just in technical matters but also in charting my path. I have also been extremely fortunate to work closely with Aaron Roth, whose door was always open and whose ability to almost-instantaneously ascertain the fundamentals of any problem has been incredibly helpful (not to mention, legendary).

I have had phenomenal teachers at Penn, including Sampath Kannan, Shivani Agarwal, Charles Epstein, George Mailath, Michael Steele, and Rakesh Vohra whose courses not only gave me the technical foundations without which I could not have succeeded in this field, but also revealed the beauty in their respective topics.

I also have been extremely fortunate to learn from a variety of researchers in various settings, including collaboration, teaching, and simply chatting about ideas, including Rachel Cummings, Ben Fish, Simon Freyaldenhoven, Vasilis Gkatzelis, Brian Lan, Jamie Morgenstern, Manolis Pountourakis, Okke Schrivers, Minchul Shin, Bo Waggoner, and Steven Wu. Some of these interactions were at several organizations – Microsoft Research, Facebook, the Federal Reserve Bank of Philadelphia

– at which I was also fortunate to get the opportunity to intern; these experiences inspired several of the problems discussed in this thesis.

Cynthia Silber and Eric Key were my parents-away-from-home. They warmly opened their home to me as a refuge when times got tough - and between home-cooked meals and the occasional math advice, I could not have made it through my program without them.

And of course, I have to thank my friends - in the same program and otherwise, in Philadelphia and beyond. Many of you are coauthors; but even those who are not, your comradery and company and friendship sustained me; from late night chats to adventures in cooking to hosting me when traveling for conferences in far away lands, the little moments of grad school will be some of my fondest memories of this time. Andrew, Alex, Ani, Ben, Brian, Chris, Darrick, Dave, Emily, Evan, Fiona, Irina, Jacob, Jason, Jinshuo, Juba, Karen, Lara, Lotem, Luai, Maher, Marcella, Marilyn, Mason, Matthew, Narayan, Omar, Paco, Pooja, Saeed, Sam, Seth, Shai, Shahin, Skanda, Sophia, Travis, Yahav, Zach - this means you. And if I have omitted others, please accept my apologies - it is only because I have spent a great deal of my cognitive capital on this dissertation, and there is little to spare.

And finally, I would be nowhere without my family. My parents, Haytham and Jumana, have provided me with endless love, support, and guidance. My mother has been my eternal editor, and I owe whatever ability I maintain to write coherently to her[1]; my father showed me the joy (and occasionally, pain) in mathematics from my earliest days. And my brother Andrew's and sister Nathalie's affection and friendship have brought me great joy and support. In their own passions, drive for excellence, and compassion, they inspire me every day.

[1]This should not be seen as an indictment of her writing.

ABSTRACT

ESSAYS ON ALGORITHMS, MARKETS, AND SOCIETY

Hadi Stephen Elzayn

Michael Kearns

This thesis examines algorithmic markets - market mechanisms with algorithms as a core component in their functioning – and markets with algorithms (that is, canonical markets for algorithmic or data-based goods and services). Our primary focus is on the analysis of these mechanisms and markets in terms of societal concerns such as fairness, privacy, and efficiency (including welfare and revenue). For algorithmic markets, we consider automated ad auctions and call auctions; for markets with algorithms, we examine both an abstract, general data-driven market from a theoretical perspective, and a specific, important data-driven market: the U.S. mortgage market. We apply a variety of theoretical tools from various subfields of Computer Science and Economics, including worst-case asymptotic runtime and sample complexity theory, the Probably Approximately Correct (PAC) Learning framework, no-regret learning algorithms, equilibrium analysis, smoothness, differential privacy, and quantitative fairness. In addition to theoretical analysis, we implement various algorithms and mechanisms and perform empirical analysis on real data in relation to the mortgage market. Our results include: new worst-case welfare guarantees, novel equilibrium characterizations, and experimental evaluation of various auction formats in the *Ad Types* setting; construction and analysis of a differentially private call auction mechanism with good performance and incentive properties; a theoretical analysis elucidating the economic forces that encourage error inequality in data-driven markets; and practical application of quantitative fairness measures to a uniquely rich and exhaustive dataset covering the mortgage market in the United States.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Algorithms are an omnipresent feature of modern life. Absent a major change in society[1], this trend will only increase. But how this trend will impact societal concerns, like the functioning of markets and the prevalence of fairness and privacy, remains unsettled. Both from an intellectual perspective as well as a practical one, this area provides a fertile ground for research. This dissertation carves out a small patch of this area, and considers several related but separate questions. Our focus is on the interaction of algorithms and markets, and how this interaction affects societal concerns.

We split this dissertation into two parts. In the first part, we study *algorithmic markets*: that is, markets that utilize algorithms in their core functioning. In the first chapter, we study automated auctions for advertising placement, which has become a many-billion dollar industry; in the second, we study double auctions, which are widely used in financial markets, and how privacy can be achieved at relatively low cost in efficiency. We dive into their details and structure to analyze their properties with respect to welfare and privacy of the participants.

In the second part, we analyze how markets that are intertwined with algorithms affect society, with a particular focus on *fairness*. The first chapter in this part

[1] See *Dune*, Frank Herbert.

(and third chapter overall) is an analysis of how economic forces create predictive inequality in machine-learning driven markets, and how competition does not, and regulation can, mitigate this. The fourth and final chapter of the dissertation focuses on one particularly important market – the U.S. mortgage market – and *empirically* documents stylized facts about fairness using tools from the machine learning and fairness literature.

We utilize and combine a variety of technical tools throughout this thesis. First, since all of our settings contain an important market component and include strategic actors, we draw on game theory and equilibrium analysis. More specialized game theoretic tools we will apply include smoothness and price of anarchy analyses, as well as classical supply and demand frameworks. Next, since we are generally concerned with algorithmic processes and learning in particular, we use both general algorithmic tools (like asymptotic runtime and sample complexity analysis) and tools from the subfield of learning theory. Specialized tools from this subject include no-regret learning algorithms, which we apply to reason about equilibrium using the connection to learning in games, and the Probably Approximately Correct (PAC)-learning framework to analyze how machine learning performance tends with sample collected, and how this translates into economic incentives. Finally, we directly apply tools from two subfields of computer science that put societal concerns front and center: the *differential privacy* literature, and the *fairness* literature. From differential privacy, we use classical mechanisms as building blocks for our own. From fairness, we leverage existing definitions and tests which we can apply empirically.

1.0.1 Overview of Results

Chapter 2: Welfare and Revenue in the Ad Types Problem In Chapter 2, we study welfare and revenue in the *Ad Types* problem, following forthcoming work. The Ad Types problem is a generalization of the standard position auction

setting, which models the interaction between a platform selling advertising slots to bidders who suffer identical decays in clickthrough rates (called *discount curves*) and thus, diminishing value, as they are placed lower in the possible set of slots. In the standard position auction setting, these discount curves are assumed to be the same for all bidders; under this assumption, algorithms for this problem can escape the complexity of a fully combinatorial auction. In the Ad Types setting, we relax this assumption, and allow different *types* of ads have different decays in clickthrough rates. However, the assumption that ads of the same *type* share the same discount curves provides an intermediate degree of generality in which it is still possible to provide guarantees.

We provide several results. First, we consider a range of auction possibilities created by product of pairing two pricing and two allocation algorithms. The pricing algorithms we consider are generalized second pricing and externality pricing, and the allocation algorithms we consider are the greedy allocation and optimal allocation. Externality pricing paired with optimal allocation gives rise to the celebrated Vickrey-Clarkes-Grove (VCG) mechanisms, which is well-known to have a weakly dominant truthful equilibrium; for each of the other formats, we characterize upper and lower bounds on welfare in the Ad Types setting. Then, we provide the first Bayes-Nash equilibrium characterization of each auction format under a very simple setting: two bidders with two slots, independent uniform valuations, and different discount types, and prove a surprising revenue equivalence result. Finally, we show that our calculated equilibria are reached experimentally when players bid in the associated auction using a no-regret learning algorithm; we then use this algorithm with simulated data to characterize the revenue of different formats in a realistic setting.

Chapter 3: Differentially Private Double Auctions In Chapter 3, we follow Diana et. al., and consider the standard double auction with uniform pricing ("call

auction") and propose a *differentially private* version. Call auctions are used in many financial markets and other settings for price discovery; in such settings, information about the willingness-to-pay of other market participants is very valuable, as it can provide an additional edge. Various ad-hoc approaches have been applied in the hope of protecting this information, but such approaches do not include any formal guarantees. Differential privacy is a mathematical notion of privacy that uses randomness to mask information with exactly such formal guarantees.

We construct a differentially private call auction by combining several mechanisms. We prove formally that our mechanism not only provides privacy but also achieves good performance (in terms of total shares cleared relative to the optimal non-private shares cleared), and show that our guarantee is tight with a nearly-matching lower bound. These guarantees make no assumptions about the particular valuations of the participants; however, we then consider agents who are assumed to behave strategically with stochastically drawn valuations. We show that our mechanism possesses good game-theoretic properties, including individual rationality and approximate incentive compatibility. Finally, we take a *learning* approach, and show that agents using variants of standard no-regret learning algorithms in a repeated version of our mechanism will learn to bid in such a way that the mechanism again clears nearly as many shares as the optimal. This result demonstrates that even if agents do not trust the mechanism's guarantee of incentive compatibility, natural approaches to attempt to profit will not affect the quality of the mechanism.

Chapter 4: Competition, Regulation, and Error Inequality in Data-Driven Markets In Chapter 4, we follow the exposition of Elzayn and Fish 2020 and turn to high-level modeling of markets in which firms produce services based on learning models from data. In particular, we investigate the economic incentives that can give rise to what we call *error inequality*: the use of machine learned mod-

els that perform worse in terms of accuracy on some groups than others. To do so, we apply core results from PAC-learning theory to take the learning side of the firms' problem seriously, and game theory and industrial organization to model the behavior of firms in response to competition and regulation.

We begin with the case of a monopolistic firm, and show that a very simple economic force – market size – drives differential investment in purchasing data for different groups. While this is not surprising from an economics-oriented point of view, this indicates that the many algorithmic innovations proposed by the fairness literature may be more band-aid than panacea. Next, we introduce a competing firm, and examine whether competition drives firms to eliminate this error inequality; we find that while competition can push firms to improve their models, the *inequality* across groups is not mitigated and may even be exacerbated. We consider various models of competition which correspond to various degrees of rationality, and under all but the most extreme, this result holds. Finally, we model government regulation as the imposition of additional constraints on the firm. We show that, depending on the form of the constraint, the regulations may or may not impose a *Price of Fairness* on the majority group relative to the unconstrained case, and characterize the degree of this Price of Fairness when it occurs.

Chapter 5: Fairness in the Mortgage Market Our final chapter, based on work conducted for the Federal Reserve Bank of Philadelphia, turns to an extremely consequential market: the U.S. mortgage market. There has historically been a great deal of racial discrimination in this market; consequently, there have also been extremely influential activist movements and policy intervention aimed at ameliorating the effects of this discrimination. We thus aim to measure *how fair* the market is today; in doing so, we apply quantitative fairness metrics developed for Fair Machine Learning to loan application and performance data obtained from public and proprietary data sets covering nearly all mortgage applications and half

of mortgages in the United States.

To obtain credible measurements in the face of potential omitted variable bias, selection effects, feedback loops, and other threats to identification, we combine a simple structural model with a focus on predictions for *marginal candidates* under various policy regimes. We find empirically that there appear to be elements of both *No Disparate Treatment* and *No Disparate Impact* regimes in practice, and we find evidence of significant disparities in approval rates, default rates, and threshold rigidity by group. We then estimate the population wide FICO-score distribution and use this to estimate a counterfactual *Pareto frontier* of possible policy regimes and their trade-offs with respect to the fairness metrics in question. Perhaps surprisingly, we find that with regard to fairness metrics, real-world measurements appear to be near the Pareto-frontier of what is achievable in the short-term using strict but differing threshold policies.

Part I

ALGORITHMIC MARKETS

Chapter 2

WELFARE AND REVENUE IN THE AD TYPES PROBLEM

2.1 Introduction [1]

In this chapter, we characterize equilibrium welfare and revenue properties of various auction formats in the *Ad Types setting*, following forthcoming work of Elzayn et al.. The Ad Types setting [37] is a generalization of the standard position auction [48, 121], which has been a workhorse in online advertising for years. In the standard position auction, there are multiple positions where the auctioneer can place ads. Advertisers care about receiving clicks on their ads, and the classical model posits a *separable* click-through-rate (CTR) model, where ad slots have an associated discount $1 \geq \delta^1 \geq \delta^2 \geq .. \geq 0$ that represents the advertiser-agnostic CTR of the slot.

The Ad Types setting [37] is a semi-separable generalization of position auctions where each ad has a publicly known type [2]—such as 'video ad', 'link-click ad' or

[1] The work in this Chapter was conducted while author was an intern at Facebook, and is based on forthcoming work with Riccardo Colini-Baldeschi, Brian Lan, and Okke Schrijvers.

[2] Type in the economics literature often refers to private information. That is not the case here: ad type refers to the conversion event that the advertiser cares about.

'impression ad'—and each ad type τ has its own associated position discount curve $1 \geq \delta_\tau^1 \geq \delta_\tau^2 \geq .. \geq 0$. All ads from the same type share the same discount curve; as such, the model represents a generalization of the position auction, while containing more structure than a general max-weight bipartite matching problem.

Colini-Baldeschi et al. [37] show that in the Ad Types setting, one can compute the optimal allocation (with respect to reported bids) and associated VCG prices using an adapted version of the Kuhn-Munkres algorithm in $O(n^2(k+\log n))$ (where n is the number of slots, and k the number of ad types). However, there are two practical considerations that need to be taken into account: First, despite the auction-theoretical benefits of VCG, in practice online advertising platforms often use a Generalized Second-Price (GSP) payment rule [8], so it is desirable to understand the impact of using GSP pricing instead of VCG. Second, in content feeds there is often a large number of ads that are allocated, making the $O(n^2(k+\log n))$ running time prohibitive, necessitating simpler non-optimal allocation algorithms.

In this chapter, we investigate what happens in the Ad Types setting when we perform the allocation using either the *greedy* or *optimal* algorithm, and run pricing using either *GSP* or *VCG* semantics. In three of the four possible combinations the resulting auction is not incentive compatible, so we investigate the revenue and welfare in equilibrium.

2.1.1 Results

We provide three sorts of results:

- **Price of Anarchy Bounds.** In Section 2.3, we provide Price of Anarchy upper and lower bounds in the Ad Types setting for all combinations of greedy or optimal allocation paired with GSP and VCG pricing. In particular, greedy allocation has an upper bound for Price of Anarchy of 4, regardless of the choice of pricing; for optimal allocation and GSP pricing, we give upper bound that depends on the bidder types and number of bidders, but not valuations.

We give lower bounds on the Price of Anarchy of 2 for greedy allocation with GSP pricing, 3/2 for greedy allocation with VCG pricing, and 4/3 for optimal allocation with GSP pricing.

- **Small Equilibrium Characterization.** In Section 2.4, we analytically characterize the existence of Bayes-Nash equilibrium in a simple case: two bidders, two slots, uniformly distributed valuations.[3] In equilibrium, the greedy allocation with GSP pricing produces and equivalent amount of revenue to the optimal allocation with VCG pricing, and that this revenue is larger than the revenue produced by either other possible mechanism (which are also equivalent to each other).

- **Evaluation on Realistic Data.** The small-equilibrium characterizations are interesting, but in order to understand if the results are representative of larger instances, we learn equilibria for bidding data from a large online advertiser in Section 2.5. We draw (normalized and anonymized) advertiser bids in various settings and equip advertisers with no-regret learning algorithms; when players use such algorithms, the empirical average of play is known to converge to *coarse correlated equilibria*. We find that for the most part equilibria on real data do not behave identically to the two bidder two slot uniform valuations case, but rather show a steeper hierarchy of revenue and welfare that conforms with intuition.

2.1.2 Related Literature

Position Auctions. Position auctions have long been the workhorse in online advertising. The seminal works of Edelman et al. [48] and Varian [121] first pro-

[3] While this may appear a very special case, explicit equilibrium characterization in auctions is notoriously complex. Most famously, in Vickrey's original paper [123] he posed an open problem to characterize the equilibrium of a two-player first-price auction with uniform valuations in $[a_1, b_1]$ and $[a_2, b_2]$, a problem that wasn't solved until nearly 50 years later [80]!

posed the separable model of the position auction—and described the generalized second-price (GSP) auction in this model—and showed that for GSP there exists an ex-post Nash equilibrium that is equivalent to the VCG outcome. Gomes and Sweeney [63] showed that GSP does not always admit a Bayes-Nash equilibrium. There is also a history of exploring alternative pricing rules for position auctions; for example Chawla and Hartline [31] study generalized first-price (GFP) semantics for position auction and show that for i.i.d. valuations the equilibrium is unique and symmetric.

Price of Anarchy and Smoothness. Since explicit equilibrium computation in auction is challenging, people have focused on Price of Anarchy bounds, i.e. using the equilibrium conditions to give bounds on the welfare in *any* equilibrium. Paes Leme and Tardos [88] were the first to give Price of Anarchy bounds for GSP. A common approach to proving Price of Anarchy bounds is to use the smoothness framework proposed by Roughgarden [111, 112], though GSP is not smooth in this sense. Lucier and Paes Leme [90] and Caragiannis et al. [26] instead show that one can use a *semi-smoothness* condition and they give almost tight Price of Anarchy bounds for GSP. Smoothness has also been applied to other payment rules, such as GFP by Syrgkanis and Tardos [117].

Complex Ad Auctions. There is a body of work that explores relaxing the separability assumption in position auctions. Our work is based on the Ad Types setting formalized by Colini-Baldeschi et al. [37]. When each ad is its own type, this model is identical to the one with arbitrary action rates that are still independent between ads, which has been studied before by Abrams et al. [2], Carvallo and Wilkens [28] and Wilkens et al. [27]. To our knowledge, no equilibrium characterizations or Price of Anarchy bounds are known in these settings. The closest is a paper by Colini-Baldeschi et al. [36] that studies the relationship between envy, regret and social welfare loss in the Ad Types setting for an alternative version of GSP called "extended GSP" using the same semi-smoothness framework as proposed by

Caragiannis et al. [26].

2.2 Model

There are n advertisers (each associated with a single ad) competing for m (ordered) slots. Each ad has a publicly known type τ_i, such as 'video ad', 'link-click ad' or 'impression ad'. Ad i of type τ_i has value-per-conversion v_i. Ads of different types have different conversion events, e.g. for a link-click ad the conversion event is a link click and for a video ad the conversion event is the user watching a video ad.

Ads in lower slots see fewer conversions, and we consider a semi-separable model[4] to capture this effect: for ads of type τ_i, we can write Pr[conversion on ad i (of type τ_i) in slot s] $= \delta^s_{\tau_i} \cdot \beta_i$ where $\delta^s_{\tau_i}$ is the slot effect for a particular ad type τ_i (e.g., the probability that a user will watch a video ad if it is shown in the sth slot) and β_i is the advertiser effect. Without loss of generality the advertiser effect has been included in the advertiser's value, i.e., if the value-per-conversion of the advertiser is v'_i, then $v_i = \beta_i \cdot v'_i$. Discount curves are *monotonically decreasing in slots*: that is, $1 \geq \delta^1_{\tau_i} \geq \delta^2_{\tau_i} \geq .. \geq 0$. In some restricted settings, we consider *geometric* discount curves that can be written as $\delta^s_\tau = c \cdot \delta^s$ for some fixed c, δ, where s is an exponent on the right hand side.

Since we consider multiple allocation and pricing formats, we write $\pi_{\mathcal{A}}(s, \mathbf{b})$ to indicate the *player* in *slot* s when $\mathbf{b}$ is the bid profile and $\mathcal{A}$ is the allocation algorithm. We will suppress the $\mathcal{A}$ when it is clear from context. We use $\sigma_{\mathcal{A}}(i, \mathbf{b})$ to indicate the *slot* that player i receives when the bid profile is $\mathbf{b}$ and the allocation algorithm is $\mathcal{A}$. We will sometimes overload notation to write $\tau(i)$ as function returning player i's ad type τ_i; this will be useful when referring not to a specific player but rather to an arbitrary occupant of a given slot. Finally, we will denote

[4]The model is semi-separable since ads of the same type share the same discount curve, but ads of different types do not.

the *optimal* allocation given a bid profile $\mathbf{v}$ with $\boldsymbol{\nu}$. That is:

$$\boldsymbol{\nu} := \operatorname*{argmax}_{\boldsymbol{\sigma} \in S_n} \sum_{i=1}^{n} \delta_{\tau(i)}^{\boldsymbol{\sigma}_i} v_i,$$

where S_n is the set of all permutations of bidders. Slightly abusing notation, we use $\nu(i)$ to denote the slot i is assigned to under $\boldsymbol{\nu}$.

Advertisers. Advertisers submit a single bid b_i for a conversion, which may or may not be their true valuation v_i. They are charged price p_i (calculated by the auction) if a conversion happens, so in expectation they are charged $\delta_{\tau(i)}^{s} p_i$. Thus, the payoff of an advertiser for a given slot at a given price is $u_i(s, p_i) = \delta_{\tau(i)}^{s}(v_i - p_i)$.

Auction Algorithms. Any auction must answer two questions: who gets what (allocation), and much how do they pay (pricing). We use $\mathcal{A} : \mathbf{b} \to \mathbf{s}$ to designate allocation algorithms, and $\mathcal{P} : \mathcal{A}, \mathbf{b} \to \mathbf{p}$ to designate pricing algorithms. Here, $\mathbf{b}$ is a vector of bids and $\mathbf{s}$ is a vector of slot assignments, so$\mathcal{A}$ maps bids to slots. Pricing algorithm $\mathcal{P}$ takes both a vector of bids and an allocation *algorithm* $\mathcal{A}$. Thus the pricing algorithm is a meta-algorithm, rather than a particular algorithm. We refer to a pair $(\mathcal{A}, \mathcal{P}_{\mathcal{A}})$ as an *auction mechanism*. In this Chapter we consider all combinations of two allocation algorithms and two pricing meta-algorithms:

- **Greedy (Allocation)** The *greedy* allocation begins with the highest slot, and among non-allocated bidders allocates the bidder whose *discounted* bid is highest (that is, $\operatorname{argmax}_{i \in \mathcal{U}_s} \delta_{\tau(i)}^{s} b_i$, where $\mathcal{U}_s$ is the set of unallocated bidders as of the time slot s is reached). For the Ad Types setting, the greedy algorithm generally does not yield the optimal allocation (see e.g. Example 1.1 in [37]).

- **Optimal (Allocation)** The *optimal* allocation computes the max-weight bipartite matching between ads and slot (where edge weights are discounted bids $\delta_{\tau(i)}^{s} b_i$), e.g. using the Kuhn-Munkres algorithm [87, 96].

- **GSP (Pricing)** The *Generalized Second Price* pricing rule executes the principle that a bidder pays the minimum bid under which they retain the slot

they were assigned to, i.e. for allocation algorithm $\mathcal{A}$ and bids $\mathbf{b}$: $[\mathcal{P}_{\mathcal{A}}(\mathbf{b})]_i := \text{argmin}_{b:\mathcal{A}(b,\mathbf{b}_{-i})_i=\mathcal{A}(\mathbf{b})_i} b$. Computing this bid is straightforward for the greedy allocation algorithm, while for the optimal algorithm we use the method of Carvallo et al [27].

- **VCG (Pricing)** The *Vickrey-Clarke-Groves* pricing rule [123, 35, 65] executes the principle that a bidder should pay their *externality*, i.e. for a allocation algorithm $\mathcal{A}$ and bids $\mathbf{b}$: $[\mathcal{P}(\mathbf{b})]_i = \sum_{j\neq i} \delta_{\tau(j)}^{\mathcal{A}(\mathbf{b}_{-i})_j} b_j - \sum_{j\neq i} \delta_{\tau(j)}^{\mathcal{A}(\mathbf{b})_j} b_j$. When $\mathcal{A}$ is the optimal allocation algorithm this yields the standard VCG algorithm. When $\mathcal{A}$ is the greedy allocation algorithm, the resulting mechanism is not incentive compatible.

Given an auction $(\mathcal{A}, \mathcal{P}_{\mathcal{A}})$, bids $\mathbf{b}$, and valuations $\mathbf{v}$, the *social welfare* is $\text{SW}(\mathcal{A}, \mathcal{P}_{\mathcal{A}}, \mathbf{b}, \mathbf{v}) = \sum_i \delta_{\tau(i)}^{\mathcal{A}(\mathbf{b})_\mathbf{i}} \cdot v_i$ and the *revenue* is $\text{Rev}(\mathcal{A}, \mathcal{P}_{\mathcal{A}}, \mathbf{b}, \mathbf{v}) = \sum_i \delta_{\tau(i)}^{\mathcal{A}(\mathbf{b})_i} \cdot \mathcal{P}_{\mathcal{A}}(\mathbf{b})_i$.

2.2.1 Solution Concepts and Learning

In this Chapter we present equilibrium results for both full-information and Bayes-Nash equilibria:

Definition 2.2.1 (Nash Equilibrium)**.** A bid profile $\mathbf{b}$ is pure strategy *Nash equilibrium* if for each player i: $u_i(\mathbf{b}) \geq u_i(b', \mathbf{b}_{-i})$ for all pure strategies b'.

Definition 2.2.2 (Bayes-Nash Equilibrium)**.** For known value distribution $\mathcal{V}$, a mapping $\mathbf{b_i}(v_i)$ for $i \in \mathcal{I}$ is a Bayes-Nash equilibrium if for player and every valuation realization v_i:

$$\text{E}_{v_{-i}\sim V_{-i}}[u_i(b_i(v_i), \mathbf{b}_{-i}(\mathbf{v}_{-i}))] \geq \text{E}_{v_{-i}\sim V_{-i}}[u_i(b', \mathbf{b}_{-i}(\mathbf{v}_{-i}))]$$

for any other mappings $\mathbf{b'_i}(v_i)$.

For an equilibrium concept, there may be multiple equilibria with different welfare. The Price of Anarchy captures the worst-case welfare compared to the optimal welfare knowing the valuations.

Definition 2.2.3 (Price of Anarchy)**.** The Price of Anarchy is

$$\max_{\mathbf{b}\in E} \frac{\sum_i \delta^{\nu(i)}_{\tau(i)} \cdot v_i}{\sum_i \delta^{\mathcal{A}(\mathbf{b})_i}_{\tau(i)} \cdot v_i},$$

where E is the set of equilibria for $(\mathcal{A}, \mathcal{P}_{\mathcal{A}})$, and recall that $\nu(i)$ is the assignment of bidder i in the optimal allocation.

2.3 Price of Anarchy

In this section, we provide characterizations of upper and lower bounds on the Price of Anarchy for (Greedy, GSP), (Greedy, VCG), and (Opt,GSP)[5]. For upper bounds on the Price of Anarchy, we leverage the *semi-smoothness* framework of [26], itself a generalization of the *smoothness* framework of [111]. For lower bounds, we construct examples of equilibria that achieve less welfare than the optimal. For results that are primarily ancillary or require involved proofs, we provide proof sketches, and defer full proofs to an expanded online version of the paper.

For Greedy GSP and Greedy VCG, we give a universal result - that is, with no requirements besides being in the Ad Types setting, and this result matches known upper and lower bounds for the position auction (though our bounds are not yet as tight). For Opt GSP, we provide *instance-optimal* bounds, where instance-optimal is with respect to the discount curves and number of slots but *universal over bidder valuations*. It is very likely that our upper bounds on the Price of Anarchy in this setting are too pessimistic; we leave improvement of these bounds to future work.

Our technique in each case will be to show that the game induced by the auction format and any valuation profile is *semismooth*, in the following sense:

Definition 2.3.1 (Semismooth [26])**.** We say that a game is (λ, μ)-semismooth if there exists some (possibly randomized) strategy (depending only on the valuation

[5]We omit (Opt, VCG) since the fact that bidding truthfully is a dominant strategy suggests alternative equilibria are unlikely to be found in practice.

	GSP	VCG
Greedy	2	3/2
Opt	4/3	NA

Table 2.1: Lower bounds on PoA.

	GSP	VCG
Greedy	4	4
Opt	$2 + (n-1)\frac{\delta^{\max}}{\delta_{\min}}$ *	NA

Table 2.2: Upper bounds on PoA. * denotes instance-optimal bounds.

of the player) such that

$$u_i(b_i', \mathbf{b}_{-i}) \geq \lambda \sum_i \delta_{\tau_i}^{\nu(i)} v_i - \mu \sum_i \delta_{\tau_i}^{\sigma(i,\mathbf{b})} v_i.$$

for all bid profiles $\mathbf{b}$.

A game can be shown to be semismooth by showing that the the following inequality holds:

$$u_i(b_i', b_i) \geq \lambda \delta_{\tau(i)}^{\nu(i)} v_i - \mu \delta_{\tau(\pi(\nu(i),b))}^{\nu(i)} v_{\pi(\nu(i),b)},$$

since if it holds, summing over players gives exactly the defining condition of semismoothness. And semismoothness directly yields Price of Anarchy bounds using the following theorem, from [26]:

Theorem 2.3.1. Suppose a game is (λ, μ)-semismooth, and social welfare is at least the sum of player utilities. Then its Price of Anarchy is upper bounded by $\frac{\mu+1}{\lambda}$.

2.3.1 Greedy Allocation Proof Recipe

A common proof structure applies to both (Greedy, GSP) and (Greedy, VCG), because of their shared allocation algorithm and the fact that both pricing algorithms,

when coupled with greedy allocation, guarantee that bidders are never overcharged. It is similar to the proof found in [26], but with additional subtlety due to disagreement in the discount factors.

To handle this subtlety, we will use the following Lemma:

Lemma 2.3.2 (Partial Monotonicity)**.** *Suppose that* $\mathbf{b}, \mathbf{b}'$ *are two bid profiles that only differ in element* i*, and* $b'_i > b_i$*. Let* σ *be the slot which* i *was assigned under* $\mathbf{b}$*. Then under greedy allocation, we have that for each slot* s *strictly above* σ*:*

$$\delta^s_{\tau(\pi(s,\mathbf{b}'))}\mathbf{b}_{\pi(s,\mathbf{b}')} \geq \delta^s_{\tau(\pi(s,\mathbf{b}))}\mathbf{b}^s_{\pi(s,\mathbf{b})}$$

Proof. First consider player i. Since i increased his bid between b to b', he achieves some slot σ' at least as high as σ.

Now, consider slots above σ'. By definition, i has not placed an effective bid higher than bidders occupying those slots (or else he would have been placed in that slot or above). So i's deviation leaves unchanged the bidder allocation and so valuations for those slots. Now, at σ', by construction, we must have that

$$\delta^{\sigma'}_{\tau(i)}b'_i \geq \delta^{\sigma'}_{\tau(\pi(\sigma',b))}b_{\pi(\sigma',b)}$$

or else i would not have been assigned to σ'. So the desired inequality holds for this slot.

Finally, consider each slot s' between σ' and σ. Notice that the set of bidders unallocated when s' is considered under b' has only changed by *losing* i and possibly *gaining* either $\pi(\sigma', b)$ or a displaced previous winners from slots between σ' and σ due to $\pi(\sigma', b)$ being displaced by i and any cascading effects. But this means that in particular $\pi(s', b)$ remains unallocated when s' is considered. Hence, if $\pi(s', b') \neq \pi(s', b)$, it can only be because the assigned bidder under b' had higher discounted value than the bidder assigned there under b. Since this holds for any s' in the range, the claim holds.

□

Now we are ready to state and prove our theorem.

Theorem 2.3.2 (Semi-Smoothness for Greedy Algorithms). Let $(\mathcal{A}, \mathcal{P}_\mathcal{A})$ be an auction mechanism. Suppose that

1. $\mathcal{A}$ is the greedy algorithm, and
2. For any bid profile $\mathbf{b}$, for every bidder we have:

$$\mathcal{P}_\mathcal{A}(\mathbf{b})_i \leq \mathbf{i}$$

Then $(\mathcal{A}, \mathcal{P}_\mathcal{A})$ is (1/2,1)-Semi Smooth.

Proof. Recall that if we can show that for any bid profile:

$$u_i(b_i', \mathbf{b}_{-i}) \geq \delta_{\tau(i)}^{\nu(i)} \frac{v_i}{2} - \delta_{\tau(\pi(\nu(i),\mathbf{b}))}^{\nu(i)} v_{\pi(\nu(i),\mathbf{b})}$$

then we will be done. So suppose b is a bid profile, and consider a deviation to bidders bidding half their value. (Notice first off that such a deviation guarantees a deviating bidder non-negative utility by Property 2.) Now, fix bidder i. For the first case, suppose that under this deviation, he receives $\mathcal{A}(b_i, b_{-i}) = \sigma' \succeq \nu(i)$ (i.e. $\nu(i)$ or better). Then we achieve the desired inequality since:

$$\begin{aligned}
u_i(b_i', \mathbf{b}_{-i}) &= \delta_{\tau(i)}^{\sigma'} v_i - \mathcal{P}_i(b_i', \mathbf{b}_{-i'}, \mathcal{A}(b_i', \mathbf{b}_{-i})) \\
&\geq \delta_{\tau(i)}^{\sigma'} v_i - \delta_{\tau(i)}^{\sigma'} \frac{v_i}{2} \\
&= \delta_{\tau(i)}^{\sigma'} \frac{v_i}{2} \\
&\geq \delta_{\tau(i)}^{\nu(i)} \frac{v_i}{2} \\
&\geq \delta_{\tau(i)}^{\nu(i)} \frac{v_i}{2} - \delta_{\tau(\pi(\nu(i),\mathbf{b}))}^{\nu(i)} v_{\pi(\nu(i),\mathbf{b})}.
\end{aligned}$$

where the 1st inequality follows by no-overcharging and the others follow by assumption or trivially.

Now suppose that instead, $\mathcal{A}(b_i', \mathbf{b}_{-i})_i = \sigma' \prec \nu(i)$. We split this into two subcases.

In the first subcase, $\frac{v_i}{2} \geq b_i$, i.e. b'_i is an upward deviation that results in i receiving σ' below $\nu(i)$. So we know that under $\mathbf{b}'$:

$$\delta^{\nu(i)}_{\tau(\pi(\nu(i),\mathbf{b}'))} b_{\pi(\nu(i),\mathbf{b}')} \geq \delta^{\nu(i)}_{\tau(i)} \frac{v_i}{2}$$

To see that this also holds when we replace $\mathbf{b}'$ with $\mathbf{b}$, turn it around. That is, we can view b_i as a downward deviation from b'_i, which cannot affect the allocation choices of any of the slots above its place before the deviation, including $\nu(i)$. But that means that the allocated bidder to $\nu(i)$ is the same under $\mathbf{b}$, so:

$$\delta^{\nu(i)}_{\tau(\pi(\nu(i),\mathbf{b}))} \mathbf{b}_{\pi(\nu(i),\mathbf{b})} \geq \delta^{\nu(i)}_{\tau(i)} \frac{v_i}{2}.$$

Then using no-overcharging, we again have that

$$\delta^{\nu(i)}_{\tau(i)} \frac{v_i}{2} - \delta^{\nu(i)}_{\tau(\pi(\nu(i),\mathbf{b}))} \mathbf{b}_{\pi(\nu(i),\mathbf{b})} \leq 0 \leq u_i(b'_i, \mathbf{b}_{-i})$$

Now, suppose that $\frac{v_i}{2} < b_i$. We know that under $\mathbf{b}'$, the bidder who gets $\nu(i)$ will have higher effective value than i, but it is not yet clear that this holds under $\mathbf{b}$. To see that this does hold, however, notice that we can view b_i as an upward deviation from b'_i. But since, by assumption, $\sigma' \prec \nu(i)$, Lemma 2.3.2 implies that in moving to $\mathbf{b}$, the values of bidders in slots above σ', which include $\nu(i)$, must increase. But then we have again that:

$$\delta^{\nu(i)}_{\tau(\pi(\nu(i),\mathbf{b}))} \mathbf{b}_{\pi(\nu(i),\mathbf{b})} \geq \delta^{\nu(i)}_{\tau(i)} \frac{v_i}{2}.$$

and the desired inequality follows as before.

□

2.3.2 Greedy GSP

Theorem 2.3.3. Let $(\mathcal{A}, \mathcal{P}_{\mathcal{A}})$ = (Greedy, GSP). Then the Price of Anarchy is at most 2.

Proof. First, by assumption, $\mathcal{A}$ is Greedy. Second, generalized second price will not charge a bidder more than their bid since under the greedy algorithm, the winner of a slot has a higher effective bid than the second bidder's bid, which is what they are charged. Hence, the conditions of Theorem 2.3.2 are satisfied, and the bound follows. □

On the other hand, we can show that the Price of Anarchy is *at least* 2.

Theorem 2.3.4. Let $(\mathcal{A}, \mathcal{P}_{\mathcal{A}})$ =(Greedy, GSP). Then the Price of Anarchy is at least 2.

Proof. Consider the following example: there are 2 slots and 2 bidders, one of type A and one of Type B. Let $\boldsymbol{\delta_A} = (1, 0)$, $\boldsymbol{\delta_B} = (1, 1)$, and let $v_A = (1 - \varepsilon)v_B$. Then the allocation (A, B) gets payoff $v_A + v_B = (2 - \varepsilon)v_B$, while the allocation (B, A) gets welfare v_B.

We claim that the following is an equilibrium: A bids 0 and B bids v_B, giving the allocation (B, A). To see that this is an equilibrium, notice that if these are the bids, $b_B > b_A$, so B will be given the first slot at a price of $b_A = 0$ for a total payoff of v_B. Since price is bounded below by 0, B could not gain by deviating any lower. On the other hand, in the second slot, A gets no value, but also is not charged, for a payoff of 0. To change anything, A would have to change the allocation, and so bid above $b_B = v_A$ - but then she would get a payoff of $v_A - v_B = (1 - \varepsilon)v_B - v_B \leq 0$; hence she also would not like to switch. And note that since $0 \leq b_A$ and $v_B \leq v_A$, neither bidder is overbidding.

But thus we see that:

$$\frac{EQ}{OPT} = \frac{v_B}{v_A + v_B} = \frac{v_B}{(2 - \varepsilon)v_B} = \frac{1}{2 - \varepsilon}$$

which can be made arbitrarily close to 1/2, and so the Price of Anarchy $:= OPT/EQ$ can be made arbitrarily close to 2.

□

Note the equilibrium described is not unique - for instance, $b_A = (1-\varepsilon)v_B$, $b_B = v_B$ would also be an equilibrium that achieves the same allocation.

For some intuition as why such a simple example can get a bad price of anarchy, notice that two slot case can be mapped to a standard second price auction for the first slot, where one bidder has a good outside option and the other doesn't. By including the outside options, a socially-minded auctioneer could do significantly better than just considering the bid and valuations of the item in question.

2.3.3 Greedy Allocation and VCG Pricing

In this section, we consider the Price of Anarchy when $(\mathcal{A}, \mathcal{P}_{\mathcal{A}})$ is (Greedy, VCG). Again, using greedy allocation guarantees the first condition of Theorem 2.3.2. It is not obvious that bidders will not be overcharged. It is, however, true, as we show in the following Lemma:

Lemma 2.3.3. *Let $(\mathcal{A}, \mathcal{P}_{\mathcal{A}})$ be the greedy algorithm with VCG pricing. We claim that for every bidder, their charge will not exceed their effective bid.*

Proof. We will prove this by strong induction. First, we relabel the bidders so that Bidder i is in Slot i post-allocation. Now, consider the removal of bidder i. First notice that this will not affect the assignment to any i' *above* i. So any price that i must pay will come from the externalities he imposes on $i' > i$.

Now, we claim that the following is true:

$$p_i = p_{j^*} + \left(\delta^i_{\tau(j^*)} - \delta^{j^*}_{\tau(j^*)}\right) b_{j^*} \tag{2.3.1}$$

where j is the bidder that is assigned to Slot i in the absence of Bidder i. (In keeping with our formal notation, $j^* := \pi(i, (\mathbf{b}_{-i}))$.)

To see that this is true, imagine re-running the auction without i included. Slots $1...i-1$ will be allocated the same way, and then at Slot i some bidder j^* will be allocated that would have been allocated further down had i been included. Now,

j^* moves up to i, he has not affected the winning bid calculations of all slots *between* i and j^* relative to what they were when i was included.

But that means that the only externalities that i imposes are those on j^* and below. Note that when we consider j^* taking the slot of i, the arrangement of the bidders below j^* will be exactly the same as if j^* were the removed bidder instead of $i*$ - but this is exactly the price that j^* pays. Hence, i's total payment is the payment of j^* plus the externality he imposes on j^*, which is $b_{j^*}(\delta^i_{\tau(j^*)} - \delta^{j^*}_{\tau(j^*)})$. But this is exactly what is claimed in Equality 2.3.1. Then we can write:

$$\begin{aligned} p_i =& p_{j^*} + b_{j^*}(\delta^i_{j^*} - \delta^{j^*}_{j^*}) \\ =& \, p_{j^*} - \delta^{j*}_{j^*} b_{j^*} + \delta^i_{j^*} b_{j^*} \end{aligned}$$

Now we invoke strong induction. Suppose that all bidders below i are not overcharged, i.e. $\forall j$ assigned to a slot below i's, $p_j \leq \delta^j_j b_j$. Then in particular, $p^*_j - \delta^{j^*}_{j^*} b_{j^*} \leq 0$, so that we conclude:

$$p_i = p_{j^*} - \delta^{j*}_{j^*} b_{j^*} + \delta^i_{j^*} b_{j^*} \leq \delta^i_{j^*} b_{j^*} \leq \delta^i_i b_i$$

where the last inequality follows by the fact that i was chosen over j for Slot i. Finally, note that Bidder n pays 0, since there are no bidders below him to exert an externality on; thus, applying strong induction starting from the bottom yields the claim.

□

Lemma 2.3.4 allows us to conclude that (Greedy, VCG) satisfies the conditions of Theorem 2.3.2, yielding the following Theorem:

Theorem 2.3.5. Let $(\mathcal{A}, \mathcal{P}_\mathcal{A}) =$ (Greedy,GSP). Then the Price of Anarchy is at most 2.

Now we consider lower bounds.

Theorem 2.3.6. Let $(\mathcal{A}, \mathcal{P}_{\mathcal{A}}) = $ (Greedy,VCG). Then there exists a conservative 3-bidder, 3-slot example with an equilibrium competitive ratio arbitrarily close to 2/3.

Proof. Let $v_A = 1 + \varepsilon$, $v_B = 1$, $v_C = 1 - \varepsilon$. Let $\boldsymbol{\delta}_A = (1, 1, 1 - 2\varepsilon)$, $\boldsymbol{\delta}_B = (1, 1, 0)$, $\boldsymbol{\delta}_C = 1, \varepsilon, \varepsilon^2$.

The welfare of (C, B, A) is $3 - 2\varepsilon - 2\varepsilon^2$, while the welfare of (A, B, C) is $2 + \varepsilon + \varepsilon^2 - \varepsilon^3$.

Suppose that each player bids their value, ie:

$$b^* = (b_A, b_B, b_C) = (v_A, v_B, v_C) = (1 + \varepsilon, 1, 1 - \varepsilon).$$

We claim this is an equilibrium and results in (A, B, C). The allocation follows since the allocation algorithm is greedy in bids. To see that this is an equilibrium, first consider what values each player is getting: A gets $1 + \varepsilon$, B gets 1, C gets $(1 - \varepsilon)\ \varepsilon^2 = \varepsilon^2 - \varepsilon^3$. With these, we can calculate what prices each player is paying: Player C pays nothing, since he is imposing no externality on A or B. B is imposing an externality on C - without B, C would get the second slot for a valuation of $\varepsilon - \varepsilon^2$ and B imposes no externality on A. So B will be charged $\varepsilon - \varepsilon^2$. Finally, A imposes the same externality on C (because without A, B would get the first slot, so C would get the second slot) and imposes no externality on B.

So the payoffs are:

$$\begin{aligned}
\pi_A(b^*) &= 1 + \varepsilon - \varepsilon + \varepsilon^2 = 1 + \varepsilon^2 \\
\pi_B(b^*) &= 1 - \varepsilon + \varepsilon^2 \\
\pi_C(b^*) &= \varepsilon^2
\end{aligned}$$

Notice that these are always positive. (The only one that could possibly be negative would be π_B, but if $\varepsilon < 1$, then $1 - \varepsilon > 0 \implies \pi_B > 0$; if $\varepsilon > 1$, then $\varepsilon^2 - \varepsilon > 0 \implies \pi_B > 0$.)

Now we consider possible deviations. Start with A. While there are an uncountable number of deviations in bid space, they are all equivalent but for their effects on A's position and price. So notice that if A were to move to second position by bidding b'_A less than b_B but more than b_C, it would receive the same payoff, because its discount rate is 1 and it imposes the same externality as before, so no such bid could improve A's payoff. If A were to bid b'_A less than b_C, it could get the third slot at a price of 0, but it would only get $1 - 2\varepsilon < 1 + \varepsilon^2 = \pi_A(b^*)$. So A has no profitable deviations. For B, improving his position cannot improve his payoff or change his externality, and moving to slot 3 would result in 0 payoff, while he currently makes positive profit. For Player C, notice that first of all, if we rule out overbidding, Player C cannot improve his position; but suppose we do not rule this out. By moving to Slot 2 (by bidding, say, $b_C = 1 + \varepsilon/2$) C would exert an externality of 1 on Player B and so get negative payoff ($1 - \varepsilon - 1 = -\varepsilon$). By moving to Slot 1 (by bidding $b_C \geq 1 + \varepsilon$) C would exert the same externality on B and so again receive negative payoff.

Hence, b^* is an equilbrium. But then we have that:

$$\frac{EQ}{OPT} = \frac{2 + \varepsilon + \varepsilon^2 - \varepsilon^3}{3 - 2\varepsilon - 2\varepsilon^2}$$

which comes arbitrarily close to 2/3 for small enough ε. □

The intuition with this example is that bidders with a high discount rate (high δ) push out the bidder with low discount rate into the tail; for either of those high-discount bidders, their *unilateral* externality is small even though, taken together, they exert a large externality. So this example is related to the fact that Nash equilibrium is about unilateral deviation, not joint deviation - there may be a better equilibrium, but the bidder losing the most cannot force a better equilibrium selection on his own.

2.3.4 Optimal Allocation and GSP Pricing

In the case of Optimal Allocation and GSP pricing, we will obtain a smoothness result that depends on the largest and smallest discounts and the number of bidders, but not on the valuation profile. The result is as follows:

Theorem 2.3.7. Suppose $(\mathcal{A}, \mathcal{P}_{\mathcal{A}})$ is optimal allocation and GSP pricing. Then the game between bidders is $(\frac{1}{2}, \frac{\delta^{\max}}{\delta^{\min}}(n-1))$-semismooth.

To prove this result, we begin by observing that GSP pricing will never charge a bidder more than his effective bid. Formally:

Lemma 2.3.4. *If bidders are conservative, then in (Opt, GSP), bid upper bounds price.*

Proof. By definition, the GSP price is the minimum the bidder could pay and earn the slot, and in particular, they could have bid exactly their bid and received their slot (because they did). Hence, the minimum they could have bid to receive the slot can never be more than whatever they actually bid. □

Now, we proceed to the proof.

Corollary 1. *The game has an instance-specific PoA of:*

$$PoA \leq 2 + (n-1)\frac{\delta^{\max}}{\delta^{\min}}.$$

If we assume that there are m slots and all discount curves are geometric and strictly ordered (e.g. $c_\tau = c_{\tau'}$ and $\delta_{\tau_1} \geq \delta_{\tau_2} \geq ... \geq \delta_{\tau_k}$ for some k, then this PoA is given by:

$$2 + 2 \cdot (n-1)\left(\frac{\delta_{\tau_1}}{\delta_{\tau_k}}\right)^m$$

We remark that this bound is potentially exponential in the number of bidders in the case of *geometric* discount curves, but linear in the case of *linear* discount curves (assuming there is a fixed set of discount curves). And while this bound is likely too pessimistic, we can give a lower bound as well:

Theorem 2.3.8. Let $(\mathcal{A}, \mathcal{P}_{\mathcal{A}}) = (\text{Opt,GSP})$. Then there exists a conservative 3-bidder 3-slot example that gets competitive ratio arbitrarily close to 3/4.

Now let's try a full information case. To do this, first we characterize what must hold in equilibrium. Then we provide examples that meet this. Again, for this we will have two bidders, two slots, two types. We will assume that in the case of a tie, A wins.

Claim 1. *Let A have discount curve* $(1, \delta_A)$*, and B have discount curve* $(1, \delta_B)$*, with* $\delta_A < \delta_B$ *(so* $\Delta := \frac{1-\delta_B}{1-\delta_A} < 1$*). Now suppose that* $\Delta^2 v_B \leq v_A \leq \Delta v_B \leq \frac{v_A}{\Delta} \leq v_B$[6]*. Then the following strategy profile is an equilibrium:*

$$\mathbf{b}^* = (\Delta(1-\delta_B)v_B + \varepsilon, \Delta(1-\delta_B)v_B)$$

for any $\varepsilon > 0$*, and for small enough* ε *neither bidder is overbidding. The auctioneer then selects* (A, B)*, but* (B, A) *would be optimal.*

Before we prove that this claim, we first show that we are not reasoning about an empty set. Consider $v_B = 1$, $v_A = \frac{1}{2}$, $\Delta = \frac{2}{3}$ (which, for example, can be obtained by $\delta_A = \frac{1}{2} < \frac{2}{3} = \delta_B$). Then $\Delta^2 v_B = \frac{4}{9} < \frac{1}{2} = v_A$, so the first inequality holds. $v_A = \frac{1}{2} \leq \frac{2}{3} = \Delta v_B$, so the second inequality holds. $\Delta v_B = \frac{2}{3} \leq \frac{3}{4} = \frac{1/2}{2/3} = \frac{v_A}{\Delta}$, so the third inequality holds, and $\frac{v_A}{\Delta} = \frac{3}{4} < 1 = v_B$ so the final inequality holds.

Notice that under this particular example, if the auctioneer selects (A, B) as claimed (and the bids truly form an equilibrium), we get a competitive ratio of:

$$\frac{EQ}{OPT} = \frac{v_A + \delta_B v_B}{v_B + \delta_A v_A} = \frac{1/2 + 2/3}{1 + 1/2 * 1/2} = \frac{7/6}{5/4} = \frac{28}{30}.$$

So we will proceed to prove the claim, and then optimize the ratio.

Proof of Claim 1. The auctioneer selects (A, B) whenever

$$b_A + \delta_B b_B \geq b_B + \delta_A b_A \iff b_A \geq \Delta b_B.$$

[6] As is always nice to check, we are not reasoning about an empty set. Consider $v_B = 1$, $v_A = \frac{1}{2}$, $\delta_A = \frac{1}{2}$, $\delta_B = \frac{2}{3}$.

But

$$b_a = \Delta(1-\delta_B)v_B + \varepsilon \geq \Delta^2(1-\delta_B)v_B = \Delta b_B$$

where the inequality follows from the fact that $\delta_A < \delta_B \implies \Delta < 1$. So the outcome is that A gets the top slot; since A will win as long as $b_A \geq \Delta b_B$, A will be charged Δb_B. B will receive the second slot, and be charged nothing. On the other hand, we note that (B,A) is optimal iff:

$$v_A + \delta_B v_B \leq v_B + \delta_A v_A \iff v_A \geq \Delta v_B.$$

This holds by assumption, so (B,A) is in fact the optimal allocation.

Now we consider possible deviations from the bid profile. For A, bidding higher does not change the allocation nor the payment, and bidding lower than its bid but more than b_B also does not affect the allocation or the payment, so the only deviation to consider is bidding less than b_B. If it does this, it will change the allocation to (B, A) and get $\delta_A v_A$ while paying nothing, but:

$$\begin{aligned} v_A - \Delta b_B &= v_A - \Delta^2(1-\delta_B)v_B \geq v_A - v_A(1-\delta_B) \\ &= v_A(1-(1-\delta_B)) = \delta_B v_A > \delta_A v_A \end{aligned}$$

where the first equality follows by the pricing rule and strategy profile, the first inequality follows from the fact that $v_A \geq \delta^2 v_B \implies -\Delta^2 v_B \geq -v_A$), and the final inequality by assumption. So deviating to be assigned the second slot would not be profitable for A.

Now consider B. Again, the only deviations that we must consider are those which change the allocation to (B, A). But if B were to deviate to such a bid, he would be charged b_A/Δ. But we have that:

$$\begin{aligned} &\frac{b_A}{\Delta} = \frac{\Delta(1-\delta_B)v_B + \varepsilon}{\Delta} = (1-\delta_B)v_B + \frac{\varepsilon}{\Delta} \\ &\implies v_B - \frac{b_A}{\Delta} = v_B - (1-\delta_B)v_B - \frac{\varepsilon}{\Delta} = \delta_B v_B - \frac{\varepsilon}{\Delta} < \delta_B v_B \end{aligned}$$

so this deviation would not be profitable for B.

Now, note that B is trivially not overbidding since $\Delta, 1-\delta_B < 1$. To show that there exists a small enough ε so that A is not overbidding, note that we need:

$$v_A - \Delta(1-\delta_B)v_B - \varepsilon \geq 0$$

so it is enough that $v_A - \Delta(1-\delta_B) > 0$. But:

$$\begin{aligned}\Delta = \frac{1-\delta_B}{1-\delta_A} > 1-\delta_B &\implies -\Delta \leq -(1-\delta_B) \\ &\implies -\Delta^2 < -\Delta(1-\delta_B) \\ &\implies -v_B\Delta^2 < -v_B\Delta(1-\delta_B)\end{aligned}$$

But then

$$v_A - v_B\Delta(1-\delta_B) > v_A - \Delta^2 v_B \geq 0$$

as desired, where the last inequality follows by assumption. Thus, we have shown that this bid profile is an equilibrium that achieves suboptimal welfare. □

Now we turn to optimizing this bound.

Theorem 2.3.9. There exists a choice v_A, v_B, $\delta_A < \delta_B$, such that the bid profile above is an equilibrium and obtains welfare arbitrarily close[7] to 3/4 of the optimal welfare. This implies that the Price of Anarchy is at least 4/3.

Proof. First, we will assume that we can find a v_A, v_B, r, with $v_A = r \cdot v_B$ for some r, such that the hypothesis of Claim 1 holds and optimize the competitive ratio over r. Then we will show that for any r, there exists a v_A, v_B such that said equilibrium holds.

Recall the hypothesis of Claim 1:

$$\Delta^2 v_B \leq v_A \leq \Delta v_B \leq \frac{v_A}{\Delta} \leq v_B$$

[7] We leave it here to avoid tie-breaking issues.

Now, letting $v_A = rv_B$ with $r < 1$, we can rewrite this as:

$$\Delta^2 \leq r \leq \Delta \leq \frac{r}{\Delta} \leq 1 \tag{2.3.2}$$

This will be the condition we will need to satisfy.

Now, let us also rewrite the competitive ratio in terms of r:

$$\frac{EQ}{OPT} = \frac{\delta_B v_B + v_A}{v_B + \delta_A v_A} = \frac{r + \delta_B}{1 + r\delta_A}$$

Notice that this function is increasing in r. So we would like to make r as small as possible while still satisfying all the hypothesis. This occurs when $r = \Delta^2$; note that Inequality 2.3.2 still holds since

$$\Delta^2 \leq \Delta^2 \leq \Delta \leq \frac{\Delta^2}{\Delta} \leq 1.$$

But if $r = \Delta^2$, then $v_A = \Delta^2 v_B$. We can then rewrite the competitive ratio as:

$$\frac{EQ}{OPT} = \frac{\delta_B v_B + \Delta^2 v_B}{v_B + \delta_A \Delta^2 v_B} = \frac{\delta_B + \Delta^2}{1 + \delta_A \Delta^2}$$

Now suppose that $\delta_A \to 0$. Then

$$\frac{EQ}{OPT} \to \frac{\delta_B + \frac{(1+\delta_B)^2}{(1-0)^2}}{1 + 0\frac{(1-\delta_B)^2}{(1-0)^2}} = \delta_B + (1 - \delta_B)^2 = 1 - \delta_B + \delta_B^2$$

This quantity is minimized at $\delta_B = \frac{1}{2}$ for a value of $\frac{3}{4}$. In that case, $\Delta \to \frac{1-\frac{1}{2}}{1-0} = \frac{1}{2}$.

The above was heuristic. More rigorously: Let $v_B = 1$, and let $\delta_B = \frac{1}{2}$. We won't fix δ_A, but rather we will assume that $\delta_A < \delta_B = \frac{1}{2}$ let it approach 0. We also set v_A as a function of δ_A: $v_A = \frac{1}{4(1-\delta_A)^2}$.

Now notice that $r = \frac{v_A}{v_B} = \frac{1}{4(1-\delta_A)^2} = \frac{(1/2)^2}{(1-\delta_A)^2} = \frac{(1-\delta_B)^2}{(1-\delta_A)^2} = \Delta^2$. Then

$$\Delta^2 \leq r \leq \Delta \leq r/\Delta \leq 1 \text{ and } \Delta^2 v_B \leq v_A \leq \Delta v_B \leq \frac{v_A}{\Delta} \leq v_B.$$

Hence, the hypothesis of Claim 1 are satisfied, so the equilibrium described is an equilibrium. Then the competitive ratio is given by:

$$\frac{EQ}{OPT} = \frac{\frac{1}{2} + \frac{1}{4(1-\delta_A)^2}}{1 + \frac{\delta_A}{4(1-\delta_A)^2}} = \frac{2(1-\delta_A)^2 + 1}{\delta_A + 4(1-\delta_A)^2}$$

Notice that at $\delta_A = 0$, this quantity is $\frac{3}{4}$, and is 1 at $\frac{1}{2}$. But notice also that the denominator, viewed independently, is a quadratic function with only complex roots, and the fraction is otherwise just a simple function of algebraic quantities. Thus, the fraction is continuous. Since it varies continuously from 1 to 3/4, it must pass through every point arbitrarily close to 3/4 from the right.

Hence, we can achieve competitive ratio arbitrarily close 3/4, so the PoA is at least 4/3. □

2.4 Equilibrium

In this section, we provide the first analytical characterization of Bayes-Nash equilibrium in the two-slot, two-bidder case with ad types and under the assumption that bidder values are drawn independently and from identical uniform distributions over the interval $[0, 1]$. In particular, we show the existence of simple equilibria that are symmetric in form and mostly natural. To find these equilibria, one may assume as a heuristic that an equilibrium exists, and derive first-order conditions; while this is a natural way to do so, ultimately, the proof is easiest when positing the existence of a linear equilibrium and verifying that the prescribed strategies are, in fact, best responses to one another. That is the approach we will take here.

For each auction type, we assume there are two slots, two discount types A and B, and one bidder of each type. We assume that the discount types have the form $(1, \delta_A)$ for type A and $(1, \delta_B)$ for type B; i.e., geometric discount curves that both have a constant factor of 1. (This assumption can be easily relaxed at the cost of carrying around some extra notation.) Throughout, we will assume without loss of generality that $\delta_A < \delta_B$, and define $\Delta := \frac{1-\delta_B}{1-\delta_A} < 1$. For the sake of efficiency, we say a bidder 'wins' if they win the first slot.

Table 2.3 displays the simple linear equilibria we discover. Notice that in each setting, player strategies are symmetric up to relabeling. In other words, the *form* of the strategy is symmetric, despite the fact that the particular strategy will differ

	GSP	VCG
Greedy	$(1-\delta_A)v_A, (1-\delta_B)v_B$	$\frac{1-\delta_A}{1-\delta_B}v_A, \frac{1-\delta_B}{1-\delta_A}v_B$
Opt	$(1-\delta_A)v_A, (1-\delta_B)v_B$	(v_A, v_B)

Table 2.3: 2 bidder, 2 type case, simple equilibrium strategies

due to different discount rates. Also, other than the VCG mechanism, each auction involves some shading. For GSP pricing, the downward shading coincides with each bidder's marginal benefit of the first slot relative to the second. But when VCG pricing is combined with greedy allocation, Bidder B shades down while Bidder A shades *up*[8].

	GSP	VCG
Greedy	$\frac{1-\delta_A}{6}\Delta^2 + \frac{1-\delta_B}{6}(3-2\Delta)$	$\frac{1-\delta_A}{6}\Delta^3 + \frac{1-\delta_B}{6}\Delta(3-2\Delta^2)$
Opt	$\frac{1-\delta_A}{6}\Delta^3 + \frac{1-\delta_B}{6}\Delta(3-2\Delta^2)$	$\frac{1-\delta_A}{6}\Delta^2 + \frac{1-\delta_B}{6}(3-2\Delta)$

Table 2.4: 2 bidder, 2 type case, equilibrium revenue

Table 2.4 gives the expected revenue for each of the equilibria described in Table 2.3. As with Table 2.3, several features are noteworthy. First, immediately we can see that both the two standard formats, as well as the two nonstandard formats, are (expected) revenue equivalent. This may be surprising, given the variation in payment rules and strategies; however, we will see that the strategies are such that the win condition and payment conditional on winning work out to be the same. Second, we note that as expected, if we allow $\delta_A = \delta_B = \delta$, we recover the equivalent revenue to the VCG mechanism for all four auction formats. This is because when discounts are the same and there are only two slots, the greedy

[8]While this may be counterintuitive, note that with greedy allocation, bidding higher increases the win probability, and under GSP pricing, bidding higher does not (directly) increase the price paid. However, overbidding results in the possibility of winning at a price higher than one's valuation.

allocation is equivalent to the optimal allocation, and GSP pricing coincides with externality pricing, so the matrix of auction formats collapses to a single row and column. Moreover, if we set $\delta = 0$, we recover the revenue of the standard second price auction with two uniform bidders (which is sensible, as if $\delta = 0$, the auction is effectively simply a second price auction for the only slot with any value). Finally, we note that it is not immediately obvious whether revenue increases or decreases with discount values (since Δ is a function of δ_A, δ_B); again, it is easy enough, if uninspiring, to take the derivative and find that revenue decreases as either discount factor increases. It may be surprising that revenue decreases when bidders can derive more total welfare, but the principle is easy to see in the extreme: if there is no difference in clickthrough rates, bidders need not bid high at all[9], as they may as well take the second slot.

These revenue results let us make *equilibrium*, rather than fixed bid[10], comparisons of revenue. In particular, simple, if involved, algebra lets us proclaim the following relationship between revenue:

Theorem 2.4.1 (Equilibrium Revenue). Consider a two-bidder, two-type, two-slot setting with bidder valuations drawn from a uniform distribution. Then in simple, linear equilibria:

$$\mathcal{R}^{\text{vcg}}_{\text{opt}} = \mathcal{R}^{\text{gsp}}_{\text{greedy}} \geq \mathcal{R}^{\text{vcg}}_{\text{greedy}} = \mathcal{R}^{\text{gsp}}_{\text{opt}}$$

Importantly, these results only apply to our simple setting; it is unclear whether the revenue, welfare, or other predictions carry over into a general setting. And indeed, in Section 2.4.6, we show that one of the least extensive generalizations does not admit such an analytically tractable characterization. While it is possible that

[9] We assume there is no reserve here; we leave as an open problem questions around designing optimal auctions with ad types.

[10] For instance, it is known in the standard position setting that GSP prices are lower bounded by VCG prices for any fixed set of bids, but such a statement makes no prediction when bidders adjust their strategies to equilibrium.

more complicated analytic equilibrium may exist and be found by clever inspection or some other method, it is difficult to foresee how such an equilibrium might be found. Moreover, it is possible that equilibrium strategies, even if they do exist, are complicated to calculate and implement. Thus, in Section 2.5, we turn our attention to empirical study of revenue under realistic bid distributions, where (coarse correlated) equilibria are *learned* via no-regret learning techniques.

2.4.1 Greedy GSP

In this setting, the higher bidder gets the top slot at a price of the lower bid, and the lower bidder gets the bottom slot at a price of 0. We obtain the following theorem:

Theorem 2.4.2. Suppose that $(\mathcal{A}, \mathcal{P}_{\mathcal{A}})$ are (Greedy, GSP). Then in the two slot, two bidder, uniform case, the strategy profile

$$(\sigma_A, \sigma_B) = ((1-\delta_A)v_A, (1-\delta_B)v_B)$$

is a Bayes-Nash equilibrium.

Proof. The easiest way to proceed is to take an ex-interim perspective; that is, consider Bidder A's perspective after she learns her valuation v_A. If A wins, she pays b_B and gets value v_A; if she loses, she gets $\delta_A v_A$ and pays nothing. Then:

$$\mathrm{E}_{v_B \sim U[0,1]}\left[u_A | b_A\right] = (v_A - \mathrm{E}[b_B | b_B < b_A])\Pr[b_B < b_A] + \delta_A v_A (1 - \Pr[b_B < b_A]) \tag{2.4.1}$$

Since we are attempting to show that $(1-\delta_A)v_A$ is a best-response to $(1-\delta_B)v_B$, we can assume that $b_B = (1-\delta_B)v_B$. Hence, A wins if and only if $v_B < b_A/(1-\delta_B)$. Under the uniform distribution, $\Pr[x < c] = c$ and $\mathrm{E}[x | x < c] = \frac{c}{2}$. Thus we can

apply these to Equation 2.4.1 to write:

$$\begin{aligned}\mathrm{E}_{v_B\sim U[0,1]}\left[u_A|b_A\right] &= \left(v_A-(1-\delta_B)\frac{b_A}{2(1-\delta_B)}\right)\frac{b_A}{1-\delta_B}+\delta_A v_A(1-\frac{b_A}{1-\delta_B}) \\ &= \frac{v_A b_A}{1-\delta_B}-\frac{b_A^2}{2(1-\delta_B)}-\frac{\delta_A b_A v_A}{1-\delta_B}+\delta_A v_A\end{aligned}$$

Taking the derivative with respect to b_A and setting to 0, we obtain the first order condition

$$\frac{v_A}{1-\delta_B}-\frac{b_A}{1-\delta_B}-\frac{\delta_A v_A}{1-\delta_B}=0 \iff b_A=v_A(1-\delta_A)$$

For completeness, note that the second derivative is $\frac{-1}{1-\delta_B}<0$, so the critical point is a maximum. Since bidding less than 0 is not a allowed and bidding more than 1 can only result in negative or zero profits, we can limit the search to the [0,1] interval. Since the payoff is continuous, the global maximum must occur at either 0, 1, or the critical point. As argued, 1 cannot be profitable, and 0 cannot give positive utility (so will always be dominated by some positive bid if $v_A>0$). Thus, the critical point is a maximum, and so A bidding $(1-\delta_A)v_A$ is best-response to the B bidding $(1-\delta_B)v_B$.

Reversing roles and considering B's perspective gives exactly the same logic. Hence, the pair of strategies form an equilibrium. □

Claim 2. *Under the linear equilibrium described above, with $\delta_A<\delta_B$, we have that the expected revenue is given by:*

$$\mathrm{E}[R]=\frac{(1-\delta_A)\Delta^2}{6}+\frac{1-\delta_B}{6}(3-2\Delta)$$

Before we prove this claim, note that if we let $\delta_A=\delta_B=0$, we immediately recover $\frac{1}{3}$, which is the revenue of the standard second price auction with two bidders drawn from $U[0,1]$. Second, if we let $\delta_A=\delta_B=\delta$, then $\Delta=1$, and we see that:

$$\mathrm{E}[R]=\frac{((1-\delta)(1)^2)}{6}+\frac{1-\delta}{6}(3-2)=\frac{1}{3}(1-\delta).$$

This has the nice interpretation that, again, with $\delta=0$ we achieve the revenue of the standard second price auction, but as $\delta\to 1$, our revenue decays linearly.

Proof. A wins if $b_A \geq b_B$, which happens when:

$$b_A \geq b_B \iff (1-\delta_A)v_A \geq (1-\delta_B)v_B \iff v_A \geq \Delta v_B$$

If A wins, she pays b_B, and so the revenue is $b_B = (1-\delta_B)v_B$; otherwise, it is $b_A = (1-\delta_A)v_A$. Thus we can write the expected revenue as:

$$\begin{aligned}
\mathrm{E}[R] &= \int_0^1 \int_0^1 R(v_A, v_B) dP(v_A) dP(v_B) \\
&= \int_0^1 \int_0^{\Delta v_B} (1-\delta_A) v_A dv_A dv_B + \int_0^1 \int_{\Delta v_B}^1 (1-\delta_B) v_B dv_A dv_B \\
&= (1-\delta_A) \int_0^1 \frac{v_A^2}{2}\bigg|_0^{\Delta v_B} + (1-\delta_B) \int_0^1 v_B v_A \bigg|_{\Delta v_B}^1 dv_B \\
&= (1-\delta_A) \int_0^1 \Delta^2 \frac{v_B^2}{2} dv_B + (1-\delta_B) \int_0^1 v_B - \Delta v_B^2 dv_B \\
&= (1-\delta_A)\Delta^2 \frac{v_B^3}{6}\bigg|_0^1 + (1-\delta_B)\left[\frac{v_B^2}{2} - \Delta\frac{v_B^3}{3}\right]\bigg|_0^1 \\
&= \frac{(1-\delta_A)\Delta^2}{6} + \frac{1-\delta_B}{6}(3-2\Delta)
\end{aligned}$$

□

2.4.2 Opt GSP

In this setting, the auctioneer chooses between the allocation (A, B) and (B, A). Note that:

$$\begin{aligned}
(A,B) \succeq (B,A) &\iff b_A + (1-\delta_B)b_B \geq b_B + (1-\delta_A)v_A \\
&\iff b_A \geq \Delta b_B
\end{aligned}$$

Suppose bidder A is the winner. Then A is charged the smallest bid b such that $b \geq \Delta b_B$, which is just Δb_B. Similarly, if B wins, he will be charged b_A/Δ.

Theorem 2.4.3. Suppose that $(\mathcal{A}, \mathcal{P}_\mathcal{A})$ are (Opt, GSP). Then in the two slot, two bidder, uniform case, the strategy profile

$$(\sigma_A, \sigma_B) = ((1-\delta_A)v_A, (1-\delta_B)v_B)$$

is a Bayes-Nash equilibrium.

Proof. We follow the same structure as the proof of Theorem 2.4.2. Consider the problem from A's perspective, and suppose that B is using a linear strategy βv_B. (In the theorem statement, $\beta = 1 - \delta_B$, and this is in fact the *only* constant that will satisfy equilibrium conditions.) Now, note that the winning condition is that:

$$(A, B) \succeq (B, A) \iff b_A \geq \Delta b_B = \Delta \beta v_B$$

If A wins the top slot with a bid b_A, then the expected payment is:

$$\begin{aligned}
\mathrm{E}[\Delta b_B | \Delta b_B \leq b_A] &= \mathrm{E}[\Delta \beta v_B | v_B \leq \frac{b_A}{\Delta \beta}] \\
&= \beta \Delta \, \mathrm{E}[v_B | v_B \leq \frac{b_A}{\Delta \beta}] = \beta \Delta \frac{b_A}{2 \Delta \beta} = \frac{b_A}{2}
\end{aligned}$$

where the second inequality follows from the properties of the uniform distribution. Then by choosing any b_A, A gets the expected payoff:

$$\begin{aligned}
\mathrm{E}[u_A(b_A | v_A)] &= \left(v_A - \frac{b_A}{2} \right) \Pr[b_B \leq b_A] + \delta_A v_A (1 - \Pr[b_B \leq b_A]) \\
&= \left(v_A - \frac{b_A}{2} \right) \frac{b_A}{\beta \Delta} + \delta_A v_A \left(1 - \frac{b_A}{\beta \Delta} \right) \\
&= \frac{v_A b_A}{\beta \Delta} - \frac{b_A^2}{2 \beta \Delta} + \delta_A v_A - \frac{b_A \delta_A v_A}{\beta \Delta}
\end{aligned}$$

Taking the derivative, the first order conditions requires that

$$\frac{v_A}{\beta \Delta} - \frac{b_A}{\beta \Delta} - \frac{b_A \delta_A v_A}{\beta \Delta} = 0 \iff b_A = (1 - \delta_A) v_A$$

as suggested. As before, it is easy to see that the payoff is concave, and that 0 and 1 are dominated strategies, so the first order condition represents a global maximum. Again, viewing this from B's perspective will give the same set of computations, mutatis mutandum, so we conclude that $((1 - \delta_A) v_A, (1 - \delta_B) v_B)$ is an equilibrium.

Notice that since the optimal b_A has no dependence on β, we can conclude that the optimal strategy responding to *any* linear strategy on the part of B is to respond with $(1 - \delta_A) v_A$. This is also true for B in response to A choosing some

linear strategy αv_A. But that means that the equilibrium we find is the *only* linear equilibrium. □

Theorem 2.4.4 (Revenue). Under the linear equilibrium described above, with $\delta_A < \delta_B$, we have that the expected revenue is given by:

$$\mathrm{E}[R] = \frac{\Delta^3(1-\delta_A)}{6} + (1-\delta_B)\Delta\left[\frac{1}{2} - \frac{\Delta^2}{3}\right]$$

Proof. Note that A wins iff $b_A > \Delta b_B$. Since $b_B = (1-\delta_B)v_B$ and $b_A = (1-\delta_A)v_A$, A wins whenever

$$v_A \geq v_B\Delta^2.$$

Now, if A wins, she pays the minimum price p such that $p \geq \Delta b_B$, which is $v_B\frac{(1-\delta_B)^2}{1-\delta_A}$. Similarly, if B wins, he pays the minimum price p such that $p \geq \frac{b_A}{\Delta} = \frac{(1-\delta_A)^2}{1-\delta_B}v_A$. So, we can write $R(v_A, v_B)$ as:

$$R(v_A, v_B) = \begin{cases} (1-\delta_B)\Delta v_B & v_A \geq \Delta^2 v_B \\ \frac{(1-\delta_A)}{\Delta}v_A & v_A \leq \Delta^2 v_B \end{cases}$$

Now we can calculate the expected revenue by again writing it as a piecewise integral:

$$\begin{aligned}
&\mathrm{E}[R(v_A, v_B)] = \\
&\int_0^1 \int_0^{\Delta^2 v_B} \frac{1-\delta_A}{\Delta} v_A dv_A dv_B + \int_0^1 \int_{\Delta^2 v_B}^1 (1-\delta_B)\Delta v_B dv_A dv_B \\
&= \frac{1-\delta_A}{\Delta}\int_0^1 \frac{v_A^2}{2}\Big|_0^{\Delta^2 v_B} dv_B + (1-\delta_B)\Delta \int_0^1 v_B v_A\Big|_{\Delta^2 v_B}^1 dv_B \\
&= \frac{1-\delta_A}{\Delta}\int_0^1 \frac{\Delta^4 v_B^2}{2} dv_B + (1-\delta_B)\Delta\int_0^1 v_B - \Delta^2 v_B^2 dv_B \\
&= \frac{1-\delta_A}{\Delta}\frac{\Delta^4}{2}\frac{v_B^3}{3}\Big|_0^1 + (1-\delta_B)\Delta\left[\frac{v_B^2}{2} - \frac{\Delta^2 v_B^3}{3}\right]\Big|_0^1 \\
&= \frac{\Delta^3(1-\delta_A)}{6} + (1-\delta_B)\Delta\left[\frac{1}{2} - \frac{\Delta^2}{3}\right] \\
&= \frac{\Delta^3(1-\delta_A)}{6} + \frac{\Delta(1-\delta_B)}{6}\left(3 - 2\Delta^2\right)
\end{aligned}$$

as claimed. □

2.4.3 Greedy VCG

Now suppose that the allocation algorithm is greedy – A wins whenever $b_A \geq b_B$ – but the pricing is VCG; that is, if A wins, she pays $(1-\delta_B)b_B$.

Claim 3. *Under greedy + VCG, the profile* $\left(\frac{(1-\delta_A)}{(1-\delta_B)}v_A, \frac{(1-\delta_B)}{(1-\delta_A)}v_B\right) = \left(\frac{v_A}{\Delta}, \Delta v_B\right)$ *is an equilbrium.*

Proof. Suppose B bids with $b_B = \Delta v_B$. Then since $b_B \leq \Delta$, any bid A makes above Δ will be equivalent in that she will certainly win and pay the same price. Thus we can write A's win probability and expected payment given winning as:

$$\Pr[b_B \leq b_A] = \begin{cases} \frac{b_A}{\Delta} & b_A \leq \Delta \\ 1 & b_A > \Delta \end{cases} \text{ and } \mathrm{E}[v_B | b_B \leq b_A] = \begin{cases} \frac{b_A}{2\Delta} & b_A \leq \Delta \\ 1 & b_A > \Delta \end{cases}$$

Hence A's payoff is:

$$\begin{aligned} u_i(b_A) &= \begin{cases} \left(v_A - (1-\delta_B)\Delta\frac{b_A}{2\Delta}\right)\frac{b_A}{\Delta} + \delta_A v_A(1-\frac{b_A}{\Delta}) & b_A \leq \Delta \\ v_A - \frac{(1-\delta_B)\Delta}{2} \end{cases} \\ &= \begin{cases} \left(v_A - \frac{1-\delta_B}{2}b_A\right)\frac{b_A}{\Delta} + \delta_A v_A(1-\frac{b_A}{\Delta}) & b_A \leq \Delta \\ v_A - \frac{\Delta(1-\delta_B)}{2} & b_A > \Delta \end{cases} \end{aligned}$$

Notice that at $b_A = \Delta$, these values coincide; beyond Δ, any value that A bids results in the same payoff. So, this payoff function is a sort of capped quadratic in b_A with the kink at Δ. Thus, to find the optimal bid, A need only compare any inner critical point with the end point (which it would even in the absence of such a kink given it were maximizing over a closed set).

On the interior section, A's first order condition is:

$$\begin{aligned} \frac{v_A}{\Delta} - \frac{(1-\delta_B)b_A}{\Delta} - \frac{\delta_A v_A}{\Delta} = 0 &\implies (1-\delta_B)b_A = v_A(1-\delta_A) \\ &\implies b_A = v_A\frac{1-\delta_A}{1-\delta_B} = \frac{v_A}{\Delta}. \end{aligned}$$

As usual, concavity gives that this is a local maximum.

But now notice that $u_i(b_A)$ is continuous up until $b_A = \Delta$, where it coincides with the next piece. Moreover, it is concave (strictly, on $[0, \Delta]$); hence, if a local maximum is reached, it must be a maximum over the interval $[0, \Delta]$, *including* the point at Δ.

So, whenever $\frac{v_A}{\Delta} \leq \Delta \iff v_A \leq \Delta^2$, it is immediate that A can do no better than bidding $b_A = v_A/\Delta$. On the other hand, if $v_A \geq \Delta^2$, then $\frac{v_A}{\Delta} \geq \Delta$. But above Δ, increasing the bid does not improve A's payoff, and so the choice of v_A/Δ prescribes a bid higher than necessary - bidding Δ would suffice. However, it also does not hurt A's payoff.

Thus, bidding $b_A = \frac{v_A}{\Delta}$ is always a best-response to B bidding $b_B = \Delta v_B$ (though it is not a unique best-response).

Now we do a similar calculation from B's perspective, supposing that $b_A = \frac{v_A}{\Delta}$. B wins if $b_A \leq b_B$ and pays $(1-\delta_A)b_A$. Again, we shall consider for the possibility of overbidding, and write the win probability and expected payoff that B will receive for any bid as:

$$\Pr[b_A \leq b_B] = \begin{cases} \Delta b_B & b_B \leq \frac{1}{\Delta} \\ 1 & b_B \geq \frac{1}{\Delta} \end{cases} \text{ and } \mathrm{E}[v_A | b_A \leq b_B] = \begin{cases} \frac{b_B \Delta}{2} & b_B \leq \frac{1}{\Delta} \\ \frac{1}{2} & b_B \geq \frac{1}{\Delta} \end{cases}$$

Then we have that

$$u_B(b_B) = \begin{cases} \left(v_B - (1-\delta_A)\frac{1}{\Delta}\frac{b_B \Delta}{2}\right)\Delta b_B + \delta_B v_B (1 - \Delta b_B) & b_B \leq \frac{1}{\Delta} \\ v_B - \frac{1-\delta_A}{2\Delta} \end{cases}$$

$$= \begin{cases} \left(v_B - (1-\delta_A)\frac{b_B}{2}\right)\Delta b_B + \delta_B v_B - \Delta \delta_B v_B b_B & b_B \leq \frac{1}{\Delta} \\ v_B - \frac{1-\delta_A}{2\Delta} & b_B \geq \frac{1}{\Delta} \end{cases}$$

Again, notice that they coincide at $b_B = \frac{1}{\Delta}$, and increasing b_B beyond $\frac{1}{\Delta}$ does not improve B's payoff. The first order condition on the interior part of the curve

is:

$$\Delta v_B - (1 - \delta_A \Delta) b_B - \Delta \delta_B v_B = 0 \implies \Delta b_B (1 - \delta_A) = \Delta v_B - \Delta \delta_B v_B$$
$$\implies b_B = \Delta v_B.$$

Again, $u_B(b_B)$ is strictly concave over $[0, \frac{1}{\Delta}]$, so this is a maximizer, and like u_A, u_B is continuous with two pieces, and the strict concavity and cap guarantees that bidding Δv_B gives at least as high payoff of bidding $\frac{1}{\Delta}$ or more. (Notice also that since $\Delta \leq 1$, the bidding strategy $b_B = \Delta v_B$ will never prescribe overbidding because $\Delta v_B \leq \frac{1}{\Delta}$.)

Thus, $b_B = \Delta v_B$ is a best response to $b_A = v_A/\Delta$, and hence the pair is a Bayes-Nash equilibrium. □

Now, we examine revenue in this equilibrium above.

Claim 4. *In this equilibrium above, revenue is given by:*

$$\mathrm{E}[R(v_A, v_B)] = \frac{\Delta^3 (1 - \delta_A)}{6} + (1 - \delta_B) \Delta \left[\frac{1}{2} - \frac{\Delta^2}{3} \right]$$

Notice that this is the *same* revenue as OPT + GSP. Why should this be? It turns out that the structure of the linear equilibrium is just so, so that the chosen allocation is the same given any two valuations *despite the allocation rules being different*, and the pricing *is also the same* despite the *pricing rules* being different.

Proof. In the equilibrium above, we have that

$$\text{A wins} \iff b_A \geq b_B \iff v_A \frac{1 - \delta_A}{1 - \delta_B} \geq v_B \frac{1 - \delta_B}{1 - \delta_A}$$
$$\iff v_A \geq v_B \Delta^2.$$

If A wins, she pays $(1 - \delta_B) b_B = (1 - \delta_B) \Delta v_B$. If B wins, he pays $(1 - \delta_A) b_A =$

$\frac{(1-\delta_A)}{\Delta} v_A$. So revenue is given by:

$$\begin{aligned}
\mathrm{E}[R(v_A, v_B)] &= \int_0^1 \int_0^{v_B \Delta^2} \frac{(1-\delta_A)}{\Delta} v_A dv_A dv_B \\
&+ \int_0^1 \int_{\Delta^2 v_B}^1 (1-\delta_B)\Delta v_B dv_A dv_B \\
&= \frac{1-\delta_A}{\Delta} \int_0^1 \frac{v_A^2}{2}\Big|_0^{\Delta^2 v_B} + (1-\delta_B)\Delta \int_0^1 v_B v_A \Big|_{\Delta^2 v_B}^1 \\
&= \frac{1-\delta_A}{\Delta} \int_0^1 \frac{\Delta^4 v_B^2}{2} dv_B + (1-\delta_B)\Delta \int_0^1 v_B - \Delta^2 v_B^2 dv_B \\
&= \frac{\Delta^3(1-\delta_A)}{2} \frac{v_B^3}{3}\Big|_0^1 + (1-\delta_B)\Delta \left[\frac{v_B^2}{2}\Big|_0^1 - \Delta^2 \frac{v_B^3}{3}\Big|_0^1 \right] \\
&= \frac{\Delta^3(1-\delta_A)}{6} + (1-\delta_B)\Delta \left[\frac{1}{2} - \frac{\Delta^2}{3} \right] \\
&= \frac{(1-\delta_A)\Delta^3}{6} + \frac{1-\delta_B}{6}\Delta \left[3 - 2\Delta^2 \right]
\end{aligned}$$

□

2.4.4 Opt VCG

Finally, recall that in the VCG mechanism, the natural equilibrium is bidders bidding truthfully. In that case, we have the following lemma:

Theorem 2.4.5. In the VCG Mechanism in this setting, revenue is given by $\frac{\Delta^2(1-\delta_A)}{6} + \frac{1-\delta_B}{6}(3-2\Delta)$.

Notice that this is, perhaps surprisingly, the same revenue as the Greedy + GSP auction. As before, this is because the winning events and conditional payments are exactly the same in this format (in this setting) as under the linear equilibrium under Greedy + GSP.

Proof. By the optimal allocation rule, A wins iff $v_A \geq \Delta v_B$, and pays $(1-\delta_B)v_B$.

Otherwise, B wins and pays $(1-\delta_A)v_A$. Hence revenue is:

$$\begin{aligned}
&\mathrm{E}[R(v_A, v_B)] \\
&= \int_0^1 \int_0^{\Delta v_B} (1-\delta_A) v_A dv_A dv_B + \int_0^1 \int_{\Delta v_B}^1 (1-\delta_B) v_B dv_A dv_B \\
&= (1-\delta_A) \int_0^1 \int_0^{\Delta v_B} v_A dv_A dv_B + (1-\delta_B) \int_0^1 \int_{\Delta v_B}^1 v_B dv_A dv_B \\
&= (1-\delta_A) \int_0^1 \frac{v_A^2}{2}\Big|_0^{\Delta v_B} dv_B + (1-\delta_B) \int_0^1 v_B v_A \Big|_{\Delta v_B}^1 \\
&= \frac{1-\delta_A}{2} \int_0^1 \Delta^2 v_B^2 dv_B + (1-\delta_B) \int_0^1 v_B - \Delta v_B^2 dv_B \\
&= \frac{\Delta^2(1-\delta_A)}{2} \frac{v_A^3}{3}\Big|_0^1 + (1-\delta_B) \left(\frac{v_B^2}{2} - \frac{\Delta v_B^3}{3}\Big|_0^1 \right) \\
&= \frac{\Delta^2(1-\delta_A)}{6} + \frac{1-\delta_B}{6}(3-2\Delta)
\end{aligned}$$

□

2.4.5 Revenue Comparison

Under the same framework, let's compare the revenue of the different auction forms. Again, this is using the revenues calculated above and provided in table 2; that is, the simple linear equilibria and assuming that $\delta_A < \delta_B$. Notice that we have shown that the VCG mechanism and Greedy Allocation + GSP pricing achieve the same revenue in this case. Ultimately this is because in both auctions, given the strategies, A wins whenever $v_A \geq \Delta v_B$ and pays $(1-\delta_B)v_B$; similarly, in both auctions, B wins whenever $v_a \leq \Delta v_B$ and pays $(1-\delta_A)v_A$.

Recall Theorem 2.4.1:

Theorem 2.4.1 (Equilibrium Revenue). Consider a two-bidder, two-type, two-slot setting with bidder valuations drawn from a uniform distribution. Then in simple, linear equilibria:

$$\mathcal{R}^{\text{vcg}}_{\text{opt}} = \mathcal{R}^{\text{gsp}}_{\text{greedy}} \geq \mathcal{R}^{\text{vcg}}_{\text{greedy}} = \mathcal{R}^{\text{gsp}}_{\text{opt}}$$

Proof of Theorem 2.4.1. The two equalities follow by inspection, so we only need to prove the inequality.

$$\begin{aligned} R^{*\text{gsp}}_{\text{greedy}} - R^{*\text{vcg}}_{\text{greedy}} =& \frac{1-\delta_A}{6}\Delta^2 + \frac{1-\delta_B}{6}(3-2\Delta) \\ & - \frac{1-\delta_A}{6}\Delta^3 - \frac{1-\delta_B}{6}\Delta(3-2\Delta^2) \\ =& \frac{1-\delta_A}{6}(\Delta^2-\Delta^3) + \frac{1-\delta_B}{6}\left(3-2\Delta-3\Delta+2\Delta^3\right) \end{aligned}$$

Expanding :

$$\begin{aligned} &= \frac{1-\delta_A}{6}\Delta^2 - \Delta^3\frac{1-\delta_A}{6} + 2\Delta^3\frac{1-\delta_B}{6} - 5\Delta\frac{1-\delta_B}{6} + \frac{3(1-\delta_B)}{6} \\ &= \frac{1-\delta_A}{6}\Delta^2 + \Delta^3\left(\frac{1-\delta_A}{6} + \frac{2(1-\delta_B)}{6}\right) - \frac{5\Delta(1-\delta_B)}{6} + \frac{3(1-\delta_B)}{6} \\ &= \frac{\Delta^2(1-\delta_A)}{6} + \Delta^3\left(\frac{3-\delta_A-2\delta_B}{6}\right) - \frac{5\Delta(1-\delta_B)}{6} + \frac{3(1-\delta_B)}{6} \\ &= \frac{\Delta^2(1-\delta_A) + \Delta^3(3-\delta_A-2\delta_B) - 5\Delta(1-\delta_B) + 3(1-\delta_B)}{6} \end{aligned}$$

Using $\delta_A \leq \delta_B \implies -\delta_A \geq -\delta_B$, we have:

$$\begin{aligned} & \frac{\Delta^2(1-\delta_A) + \Delta^3(3-\delta_A-2\delta_B) - 5\Delta(1-\delta_B) + 3(1-\delta_B)}{6} \\ & \geq \frac{\Delta^2(1-\delta_B) + \Delta^3(3-3\delta_B) - 5\Delta(1-\delta_B) + 3(1-\delta_B)}{6} \\ & = \frac{1-\delta_B}{6}\left[\Delta^2 + 3\Delta^3 - 5\Delta + 3\right] \end{aligned}$$

On the range $\Delta \in [0,1]$, the inner function is positive. To see this, one can either graph the function using a computer algebra system, or prove this analytically. For completeness: Note that $3+\Delta^2+3\Delta^3-5\Delta$ is bounded below by $3+\Delta^2+\Delta^3-5\Delta$. So it suffices to show that the latter is positive on $\Delta \in [0,1]$. So let $f(\Delta) = \Delta^2 + \Delta^3 - 5\Delta$. Then notice that $f(0) = 0$, $f(1) = -3$, and $f'(\Delta)$ is given by $3\Delta^2 + 2\Delta - 5$. Since $\Delta < 1$, f' is always negative on $[0,1]$. But that means that, given that $f(0) = 0$ and $f(1) = -3$, f cannot go below -3 on the interval (otherwise it would have to have a positive derivative at some point to come back up to -3).

Hence, we conclude that $f(\Delta) \geq -3\ \forall \Delta \in [0,1]$, and so $3 + f(\Delta) \geq 0$. Tracing back through the inequalities, this gives $R^{*\text{gsp}}_{\text{greedy}} \geq R^{*\text{vcg}}_{\text{greedy}}$, and the claim follows. □

2.4.6 More complicated settings

Unfortunately, though the two bidder case admits elegant linear equilibria, complicating the setup (even to as simply as two slots, two bidders of one type and one bidder of another) immediately eliminates hope of finding a simple linear equilibrium in general. To see this, one can posit a linear equilibrium again symmetric up to discount types. Then beginning with the rare player, one can attempt to solve for this linear equilibrium, and one way to attack this is to split the outcome of the game into two stages: first, there is a preliminary elimination from contention based on player bids, and then the bids of the remaining players are applied in an auction. Using this compound game structure, one can easily write the payoff of a bid, and then it is straightforward to show that if the common type is playing any linear strategy, the best-response for the rare player is *not* linear. Hence, there is no equilibrium where players of the same type play symmetrically and all strategies are linear.

2.5 Experimental Results

In this section, we run two experiments. Both leverage the Exponential Weights algorithm, described in [1]. In the first experiment, we use no-regret learning algorithms in the two-slot, two-bidder setting (following a technique described in [68]) to *learn* equilibria; we find that these learned equilibria closely match our predicted Bayes-Nash equilibria. In the second, we evaluate equilibrium revenue and welfare in a 10-bidder, 5-slot setting with valuations drawn from simulated distributions, where again we use no-regret learning to learn the equilibria (which would be far too complicated to recover analytically). We find that in practice, revenue seems to exhibit a strong hierarchy across auction types, but welfare does not. We also note that the strength of the revenue hierarchy and absolute level of revenues are larger when valuations are independent rather than correlated.

ALGORITHM 1: Generic Exponential Weights algorithm.

Input: Learning rate η. Bidspace $b^1, ...b^k$, Auction interface $(\mathcal{A}, \mathcal{P}_\mathcal{A})$, opponent bids b_{-i}.

Set $\mathbf{w}_t(b) \leftarrow \frac{1}{k}$ for $b = b^1, \ldots, b^k$ ▷ Initialize uniform weights.

for $t \in 1...T$ **do**

$b_t \sim \mathbf{w}_t$ ▷ Draw bid from distribution

$u_t(b^j) \leftarrow u(\mathcal{A}, \mathcal{P}_\mathcal{A}, b^j, b_{-i})$ ▷ Observe payoff of each bid b^j given auction and fixed opponent bids.

$\mathbf{w}_{t+1}(b^j) \leftarrow \frac{\exp(\eta u_t(b^j))}{\sum_\ell \mathbf{w}_t(b^\ell)\exp(\eta u_t(b^\ell))} \cdot \mathbf{w}_t(b^j)$ for $k = 1, \ldots, \mathbf{v}_j^b$ ▷ Update the weights.

end

2.5.1 Validating Theoretical Equilibria

In our first experiment, we modify the technique described in [68] to *learn* equilibria in the two-bidder two-slot uniform distribution setting – that is, the setting for which we obtain analytical results in Section 2.4. The purpose of this, besides curiosity, is twofold. Narrowly, the fact that it agrees so precisely with our theoretical results both validates the theory and heartens our confidence[11] in our implementation[12] of the mechanisms described. But more broadly, it suggests that the approach of *learning* equilibrium, which may be the only feasible approach given the computational intractability, may in fact also be the "right" approach in terms of producing reasonable results in more general settings.

The idea of this approach is simple. It is well-known that if agents in a repeated game use no-regret learning algorithms to decide their actions, then the empirical time average of play converges to coarse correlated equilibrium. What [68] shows is that we can use the *population interpretation* of a Bayesian game to extend these

[11] The code performs as expected on a reasonably wide variety of unit tests. But more validation is always nice.

[12] As of this writing, our implementation has not been open-sourced due to it being written as part of a summer internship; we hope to obtain approval for public release at some point in the near future.

results analogously for games of incomplete information.

The population interpretation views a game of incomplete information between n players with t types as instead a game between nt players, n of which are selected to play each round (where the selection is of one type from each original player). [68] suggests that each *type* be equipped with its own no-regret learning algorithm; then, results on learning coarse correlated equilibria can be extended to learning Bayesian coarse correlated equilibria. Of course, this approach assumes finite types and actions; we will approximate this by discretizing both the space of bids and space of valuations.

We use this method as our basis. However, we modify it by introducing an exploration period. This is inspired by [55], which shows that if a long enough period of pure exploration is given, then contextual no-regret learning algorithms in repeated auctions will converge to the *specific* natural equilibria that one might hope to find, e.g. truthful bidding in the second price auction or the standard shading in a first-price Bayesian auction. This is in stark contrast to what can be proven *without* an exploration period. (See, e.g. [54].)

Our protocol is thus given in Algorithm 2. In words, we simply discretize the value space into values and discretize the bid space[13]. We instantiate a copy of the Exponential Weights algorithm for bidders for each valuation and each type.

For specific parameter values, we pick $\delta_A = 0.37$ and $\delta_B = 0.55$; these are (rounded) discounts such that $\Delta^2 = \frac{1}{2}$. As in the theoretical section, we assume that discounts for the first slot are 1, and we approximate uniformly distributed valuations over $[0, 1]$ with valuations in increments of 0.1 and uniform weight. Our bid space is discretized separately for each player (i.e. for each valuation) to span 21 equally spaced increments between 0 to twice the valuation[14]. We run for 500,000

[13] We include bids above 1 to allow players to overbid, which is prescribed in the VCG-Greedy equilibrium.

[14] I.e. for a bidder with valuation 1, the bid choices are 0,0.1,0.2,...1.1,1.2...2, but for bidder with valuation 0.5 they are $0, 0.05, ...0.5, ...1$.

rounds of learning after a pure exploration period of 10,736 rounds[15]. Then we calculate the mean bid for each valuation over the period after exploration. The results are visible in Figure 2.1.

ALGORITHM 2: Experiment 1 Protocol

Input: Value Discrization d_v, Bid Discretization d_b, Number of exploration rounds T', Number of rounds T, δ_A, δ_B,

Output:

for $(\mathcal{A}, \mathcal{P}_\mathcal{A}) \in \{[Greedy, Opt] \times ([GSP, VCG])\}$ **do**

  for $v \in \{0, \frac{1}{d_v}, \frac{2}{d_v}, ..., 1\}$ **do**

    Initialize bidder of type A with value v, discount factor δ_A, and exponential weights with action space $\{0, \frac{1}{d_b}, \frac{2}{d_b}, ...1, 1 + \frac{1}{d_b}, ...2\}$

    Initialize bidder of type B with value v, discount factor δ_B and exponential weights with action space $\{0, 1/d_b, 2/d_b, ...1, 1 + 1/d_b, ...2\}$

  end

  for $t \in 1, 2..., T$ **do**

    Select random bidder of type A bidders and random bidder from type B bidders. **if** $t > T'$ **then**

      Run auction with $(\mathcal{A}, \mathcal{P}_\mathcal{A})$ and the given bidders with their bid distributions.

    end

    else

      Run auction with $(\mathcal{A}, \mathcal{P}_\mathcal{A})$ and the given bidders but bid randomly.

    end

    Update ExpWeights for selected players.

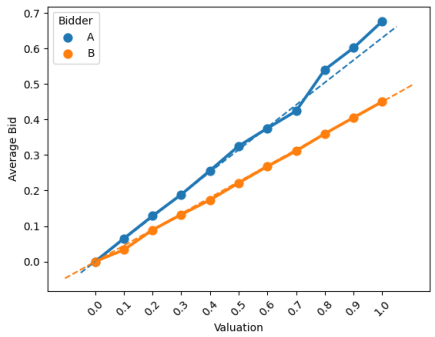
  end

end

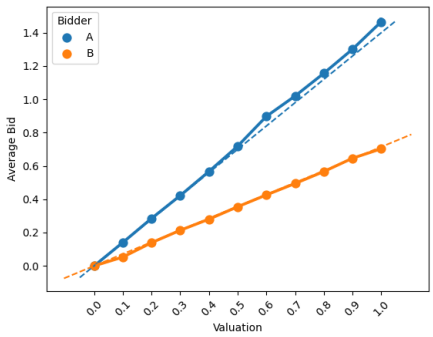
Qualitatively, the figures match our theoretical predictions quite well, with deviations explainable by error due to discretization and incomplete learning. In particular, for this choice of δ_A and δ_B, we should expect the strategies to be $(0.63v_A, 0.45v_B)$ in GSP-Greedy and GSP-Opt, (v_A, v_B) in VCG-Opt, and $(1.4v_A, 0.7v_B)$

[15] These are calculated to be in line with the requirements of [55], but because of differences in our setting relative to theirs, this calculation should be considered very rough.

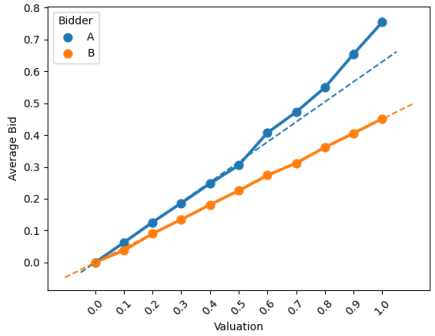
in VCG-Greedy. (The latter may be particularly surprising as it calls for overbidding on A's part.) Each of these equilibria are apparent in the figures. The discrepancy between predicted and observed results is more apparent at higher valuations. This is likely due to our method of discretization: in using a constant number of actions across valuations, the gap between potential actions is varying, and thus the approximation to a continuous actionspace is less precise for higher valuations.



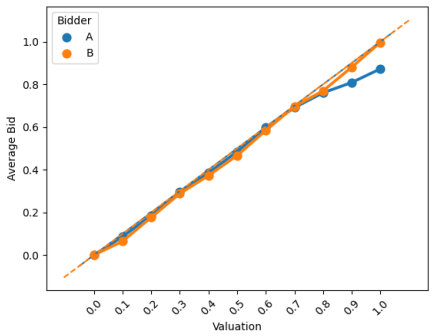
(a) GSP-Greedy



(b) VCG-Greedy



(c) GSP-Opt



(d) VCG-Opt

Figure 2.1: Average bids by value after exploration period. Dotted lines indicate theoretical predictions.

2.5.2 Experiment 2

In our second experiment, we evaluate welfare and revenue obtained by the various mechanisms in simulated data. We consider a 10-bidder, 5-slot setting, which is

rather large among empirical studies of these mechanisms (but still small relative to the auctions seen in practice). We generate two synthetic datasets. In our first dataset, we fix independent lognormal distributions (with the underlying normal distributions having standard unit variance but varying means) for each bidder and draw values independently to use as undiscounted bidder click valuations. In our second, we similarly draw values from lognormal distributions, but induce correlation of click valuations *across bidders* by drawing common component drawn from a uniform distribution, and using a simple average of the independent and common components for each bidders click valuation. We give parameters for valuations in Table 2.5.

Bidder	1	2	3	4	5	6	7	8	9	10
Quantity										
Underlying Mean	-2	-1	-0.5	-1.2	-3	-1.7	-0.9	-0.5	-1.2	-3.5
Constant	1	1	1	1	1	1	1	1	1	1
Factor	0.9	0.9	0.9	0.9	0.8	0.8	0.7	0.7	0.6	0.6

Table 2.5: Simulated data parameters for Experiment 2, including means of underlying normal distribution and constant and factor for geometric discount curves.

A protocol for a single round is as follows. We draw a number to indicate the row of the dataset to use and set bidder click valuations according to that row. (Choosing the same row ensures that the bidders will have a common component in the correlated dataset, and does not affect independence in the independent dataset.) Then, for 1000 rounds, bidders play a repeated auction game where their valuations are all fixed as of the initial draw, and players use exponential weights to play. That is, each round bidders maintain a current distribution over actions, draw an action, and then, based on the outcome of the round and what they would

have received had they played every other action[16], they update their weights[17]. We then sample a strategy profile from the time-averaged joint distribution by uniformly selecting a time period and drawing bidders' bids from the exponential weights distributions of that given round; we take 100 such samples and run the auctions with these bids. Then we average these together to obtain an estimate of the revenue of each mechanism for the initial valuation draw. And for each each auction format, we repeat this entire process for 100 total bidder valuation draws.

We plot revenue in Figure 2.2a and welfare in Figure 2.3a.

Results Our experimental results highlight several important qualitative conclusions. First, the average revenues achieved in Figure 2.2 suggest a clear revenue hierarchy: GSP pricing obtains significantly higher revenue than VCG pricing, and greedy allocation seems to obtain, at least in the correlated case, somewhat more revenue than optimal allocation. Notice that this hierarchy is quite different than Theorem 2.4.1, and matches the intuition that GSP generally prioritizes revenue rather than incentive compatibility as in VCG. Second, revenue obtained under independent valuations is significantly higher than under correlated valuations; interestingly, this difference appears larger for GSP pricing than VCG. Finally, note that in Figure 2.3, we see that welfare too is much higher under independent valuations than correlated ones, and that there does not appear to be much difference in welfare across auction formats. One potential explanation may be that the bulk of welfare is being contributed by high-value but steep-discount bidders, and greedy and optimal allocations will treat such bidders similarly, at least in the extreme case.

[16]We evaluate every unplayed bid choice by re-running the auction with all *other* players' bids fixed and calculating the counterfactual payoffs.

[17]Before they update, we store each round's distribution separately for later use.

ALGORITHM 3: Experiment 2 Protocol

Input:

Output:

for $(\mathcal{A}, \mathcal{P}_{\mathcal{A}}) \in \{[Greedy, Opt] \times ([GSP, VCG])\}$ **do**

for $s \in 1, 2..., 100$ **do**

Draw 10 valuations. Initialize bidders with values and fresh Exp. Weights.

for $t \in 1, 2, ..., 1000$ **do**

Initialize and run a 5 slot auction using $(\mathcal{A}, \mathcal{P}_{\mathcal{A}})$ Draw bids and fix them.

for $i \in \mathcal{I}$ **do**

for $b' \in \{\frac{1}{d}, \frac{2}{d}, ..., 1\}$ **do**

Re-run auction with all other players' bids fixed, but player i using b'. Save payoff.

end

end

Update ExpWeights.

end

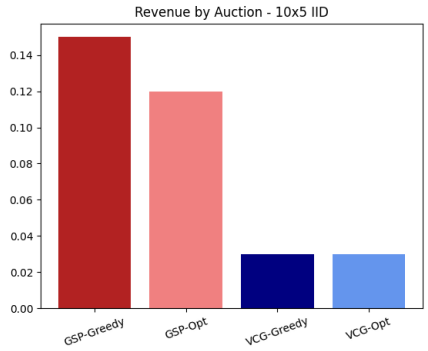
for $t' \in 1, 2, ...100$ **do**

Draw number uniformly at random from the total number of learning steps to become round number. Draw bid for each bidder from time average of ExpWeights at that round. Run auction with these bids Save round info.

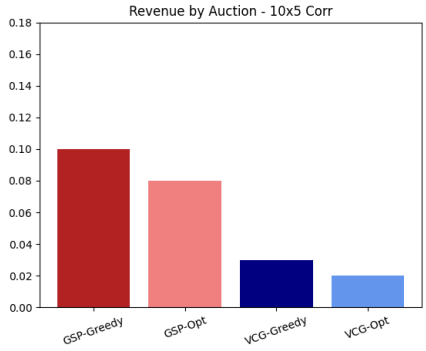
end

end

end

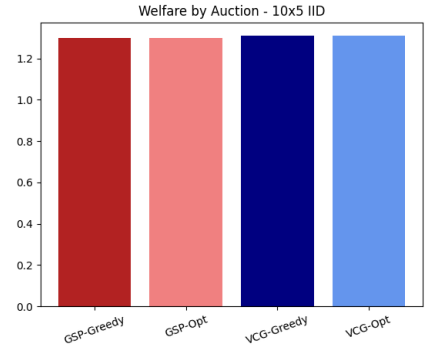


(a) Revenue by auction format for 10 independent bidders and 5 slots.

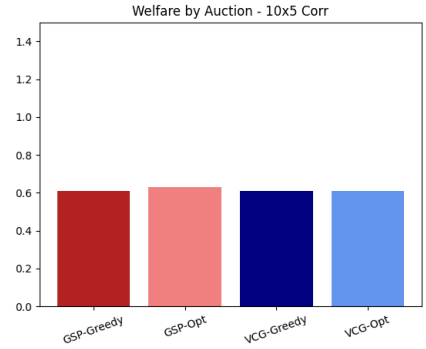


(b) Revenue by auction format for 10 correlated bidders and 5 slots.

Figure 2.2: Revenue by Auction



(a) Welfare by auction format for 10 independent bidders and 5 slots.



(b) Welfare by auction format for 10 correlated bidders and 5 slots.

Figure 2.3: Welfare by Auction

Chapter 3

DIFFERENTIALLY PRIVATE DOUBLE AUCTIONS

3.1 Introduction [1]

In modern financial markets, massive resources are directed towards what can be considered ad-hoc privacy mechanisms, intended to allow participants to cloak their trading activity and intentions. Such efforts occur both in the exchanges themselves and in the algorithmic trading services offered by large brokerages. In this work, we provide a differentially private (DP) version of classical one-shot double auctions (also known as "call auctions"). Frequent instances of DP call auctions could potentially simplify the convoluted efforts at providing trading secrecy that are rampant in today's markets while still permitting dynamic price discovery.

Current electronic exchanges offer a staggering variety of order types and mechanisms meant to provide specific types of privacy. Dark pools were introduced to allow large-volume counterparties to discover each other away from the so-called "lit" markets where high-frequency traders (HFTs) are prevalent. Order types re-

[1] This Chapter is based on joint work [40] with Emily Diana, Michael Kearns, Aaron Roth, Saeed Sharifi-Malvajerdi, and Juba Ziani.

stricting execution with small-volume counterparties are meant to provide similar protections. Hidden and "iceberg" orders in the lit exchanges provide secrecy at the expense of time priority in the standard continuous limit order book. The relatively new exchange IEX was created to foil the latency arbitrage of HFT by introducing a "speed bump" for all incoming orders. On the brokerage side, algorithms executing large client trades attempt to minimize visibility by breaking orders up over time and across exchanges and employ randomization in both timing and sizing to avoid detectable "heartbeats."

These efforts are all ad-hoc in the sense that they each protect market participants from rather specific forms of detection or exploitation. While well-intentioned, they have contributed significantly to the complexity of modern electronic markets. At the same time, it is also widely understood that there are limits to the privacy that can be provided for large trades executed in short periods, and there is a large academic and practical literature on theories of market impact (see [61, 62] for an overview) and algorithms for minimizing it. This literature identifies a trade's *participation rate* — the ratio of its volume to that of the overall market during the trade's execution — as the key determinant of market impact.

Our main conceptual contribution is the development of DP call auctions as a mechanism providing privacy against *all* forms of attack or detection, up to the participation rate of a trade. In this formulation, we provide a per-share privacy guarantee determined by the sensitivity of the call auction, which in turn determines the amount of noise added. Trades with higher participation rates will unavoidably have less privacy than those with smaller ones, but the nature of the privacy will now be as general as possible. Repeated DP call auctions also enjoy graceful degradation of the privacy guarantee. Furthermore, we can (informally) relate our results to standard market impact theories via the shared notion of participation rate and show that, under natural conditions, DP call auctions clear a near-optimal number of shares under the predictions of the "square root law" of market impact. We ana-

lyze our DP mechanisms extensively, including its incentive properties and behavior under natural no-regret learning dynamics by market participants.

We note that (non-private) call auctions are already common in modern markets. In particular, both NYSE and NASDAQ hold call auctions (also sometimes called "crosses") to establish opening and closing prices in U.S. equities [99, 98]; in the Tokyo Stock Exchange there are additional intraday call auctions, which are also the subject of academic study (e.g. [30, 29]). The influential paper [24] (discussed at greater length in Related Work below) proposes and analyzes frequent intraday (again non-private) call auctions specifically as a defense against latency arbitrage; see also [124]. Our work can be seen as a continuation of this line of thinking, in which frequent intraday DP call auctions could provide even more general privacy guarantees to all market participants.

Outline and Summary of Results: At a high level, our results fall into three broad categories:

1. **The development and analysis of (jointly) differentially private call auctions**. We carry this out in Section 3.3. We initially present this purely as an algorithm design task, abstracting away incentive properties. We prove bounds relating the *privacy* properties of the mechanism, the *number* of shares it is guaranteed to clear compared to the optimal benchmark, and the net *inventory* that the mechanisms may have to take on. (Unavoidably, jointly differentially private call auctions cannot exactly match the number of buyers and sellers and so will have to take on a net position of shares itself to clear the market — we prove that this net position is small.). We also prove a lower bound showing that our mechanisms are near optimal amongst all differentially private mechanisms. We explore the connection between our guarantees and theories of market impact in Section 3.3.5.

2. **The analysis of incentive properties and learning dynamics**. Having

developed our algorithms, we turn our attention to how buyers and sellers should interact with them. First, in Section 3.4, we show that our algorithm is ex-post individually rational and approximately dominant strategy truthful for agents who wish to trade only a small number of shares, with a guarantee that degrades gracefully in the size of the desired trade. (We note that this is a *stronger* incentive guarantee than standard non-private call auctions.) We then study the global behavior that results when agents interact with a repeated version of one of our mechanisms using *learning dynamics*: we show that although an abstract guarantee of no-regret learning is not enough to guarantee convergence to the optimal number of trades, a small modification of the exponential weights learning algorithm (informally, a modification that still guarantees the no-regret property, but breaks ties in favor of trading whenever such ties exist) does converge to the optimal number of trades.

3. **Simulation Results**. Finally, in Section 3.5, we conduct simulations in both *one-shot* and *repeated* settings, showing that in the settings considered, the realized outcomes of our mechanisms tend to be significantly better than the worst-case guarantees of our theorems.

Related Work: Our work relates to several large strands of literature. Prominently, the study of double auctions dates back to the early days of mathematical economics. [101] provides an introduction to double auctions, and a useful survey from a computer science perspective can be found in [102]. Our modeling of the strategic framework in which agents participate in the double auction is broadly consistent with this literature.

Of particular note are [24] and [124], which both propose frequent call auctions to eliminate latency arbitrage.[2] The work of [24] first establishes the *empirical*

[2] *Latency arbitrage* is the opportunity for traders to simultaneously buy and sell nearly or exactly identical securities on different exchanges (e.g. Chicago's Mercantile Exchange and the

availability of latency arbitrage opportunities for even highly traded securities, and shows moreover that competition between traders has not eliminated this opportunity over time. Instead, it has resulted in an "arms race" for speed, with arbitrage windows becoming shorter over time, but arbitrage profit per unit remaining essentially constant. The authors then propose a solution to mitigate latency arbitrage: repeated high-frequency call auctions. Using a game theoretic approach, they model how the "sniping" process results in arbitrage opportunities in the continuous limit order book; in their model, the profit opportunity (along with arms race) is an equilibrium constant, even despite improving technology. Then, using the same underlying model of firm behavior, the authors show that repeated call auctions eliminate these arbitrage opportunities and cause firms not to choose to invest in speed, ending the arms race. We follow in the spirit of [24], but note that their solution does not mitigate the problem of privacy, and in particular does not solve the issue of the proliferation of ad-hoc and increasingly complex trading algorithms. (The earlier work of [124] performs extensive simulation studies that establish the salutary effects of frequent call auctions on latency arbitrage.)

Our work is connected to, and leverages tools from, the broad literature on differential privacy introduced by [45]; for an overview, see, e.g. [46]. The most related strand of this literature is the connection between differential privacy and mechanism design, first made by [94]. In particular, they observed that differentially private mechanisms inherit strong incentive properties. For many mechanism design tasks that involve the allocation of a resource to individuals, it is not possible to satisfy differential privacy in the standard sense over allocations: in cases like this, the relevant solution concept is *joint differential privacy* [83]. This solution concept has been used in a number of mechanism design settings, including max-welfare matchings and other allocation problems [72, 73], stable matchings [79],

New York Stock Exchange) in the instant where price has changed on one exchange but remains "stale" on the other; it is described in the popular book *Flash Boys* [89].

equilibrium selection problems [109, 39], and tolling problems [108]. In particular, although joint differential privacy can be used as a tool to achieve truthfulness, not all jointly private mechanisms are approximately truthful, and more specialized arguments are needed. Finally, while [33] have shown how to privately compute near-optimal prices in double auctions, their process does not guarantee end-to-end joint differentially privacy when taking trade allocations into account, unlike this work.

3.2 Model and Preliminaries

3.2.1 Model

We consider a call auction setting with n^s sellers and n^b buyers; we let $\mathcal{S}$ be the set of sellers, $\mathcal{B}$ be the set of buyers, and $n = n^s + n^b$. Each seller $i \in \mathcal{S}$ has one unit of a security for which it has a *value* v_i^s; each buyer $j \in \mathcal{B}$ wishes to purchase one unit of the security for which it has a value of v_j^b. We let $\mathbf{v}^s = (\mathbf{v}_1^s, \ldots, \mathbf{v}_{n^s}^s)$ be the vector of all sellers' valuations and $\mathbf{v}^b = (\mathbf{v}_1^b, \ldots, \mathbf{v}_{n^b}^b)$ the vector of all buyers' valuations. We assume valuations are drawn from a discrete set P; without loss of generality, we let $P = \{1, 2, \ldots, V\}$ for some integer V.

Agents report their valuations in P directly to a mechanism $\mathcal{M}$.[3] Based on the agents' reports, the mechanism selects a clearing price $p \in P$ and an allocation vector $\mathbf{a} = (\mathbf{a}^s, \mathbf{a}^b)$, where $\mathbf{a}_i^s$ (resp. $\mathbf{a}_j^b$) is equal to 1 if seller i (resp. buyer j) is selected to participate in a trade and 0 otherwise. The mechanism concludes by buying a share at price p from every seller i with $\mathbf{a}_i^s = 1$ and selling a share at price p to every buyer with $\mathbf{a}_j^b = 1$.[4]

[3] We will argue in Section 3.4 that it is in every agent's best interest to report his valuation to the mechanism truthfully, hence our mechanisms can work with the agents' valuations without loss of generality.

[4] In principle, mechanisms can choose *non-uniform* pricing; that is, different agents could be

Privacy Constraints The outcomes of the mechanisms we consider are functions of the agents' reports, which themselves depend on their valuations. In turn, these outcomes may leak information about the participants' valuations. This provides motivation for designing call auctions that protect the privacy of the participants. In this Chapter, we do so using *differential privacy* ([45]). We will design our mechanisms to release the clearing price p in a differentially private fashion, and the allocation vector $\mathbf{a} = (\mathbf{a}^s, \mathbf{a}^b)$ in a *jointly* differentially private manner [83]. Differential privacy and joint differential privacy are formally defined in Section 3.2.2.

Mechanism Designer's Objective The main objective of our mechanisms for call auctions is to maximize the volume of trades between buyers and sellers. However, because of the randomization that we will need to add to achieve differential privacy, our mechanism will inevitably incur several kinds of cost. First, the *payoff* of the mechanism, given by the number of shares cleared, will generally be lower than the optimal payoff that could have been reached absent differential privacy. Second, we will have to deal with situations in which the number of sellers and the number of buyers who are selected to trade differ because of the noise added to the allocation rule for privacy concerns; this creates an *inventory* in which some of the trades must be fulfilled by the mechanism itself (when there are more sellers selected than buyers, the mechanism buys surplus shares from the sellers; when there are more buyers selected than sellers, the mechanism sells to the buyers from its own reserve of shares). We will aim to keep the inventory of our private auction mechanisms as small as possible. Formally, the payoff and the inventory of a mechanism $\mathcal{M}$ are defined as follows:

Definition 3.2.1 (Payoff and Inventory of a Mechanism $\mathcal{M}$)**.** For any mechanism $\mathcal{M}$ outputting a price p and an allocation vector $\mathbf{a}$, the payoff is the number of

charged different prices based on their reports. Here, we only consider *uniform* pricing mechanisms, as is commonplace in the double auction literature.

shares cleared by $\mathcal{M}$:

$$\Pi(\mathcal{M}) = \min\left\{\sum_{i\in\mathcal{S}} \mathbf{a}_i^s, \sum_{j\in\mathcal{B}} \mathbf{a}_j^b\right\}$$

, and the inventory of $\mathcal{M}$ is the number of allocations that must be fulfilled by the mechanism:

$$I(\mathcal{M}) = \left|\sum_{i\in\mathcal{S}} \mathbf{a}_i^s - \sum_{j\in\mathcal{B}} \mathbf{a}_j^b\right|$$

.

The main benchmark we use to measure the performance of our mechanisms is the maximum number of trades that can be obtained (absent differential privacy) while setting a uniform price p and guaranteeing every agent non-negative utility.[5] Formally, our benchmark is given by[6]

$$\text{opt} = \max_{p\in P} \min\left\{\sum_{i\in\mathcal{S}} \mathbf{1}\left[\mathbf{v}_i^s \le p\right], \sum_{j\in\mathcal{B}} \mathbf{1}\left[\mathbf{v}_j^b \ge p\right]\right\}. \tag{3.2.1}$$

3.2.2 Differential Privacy

Let $\mathcal{D}$ be a *data universe* from which a data set D of size n is drawn. In the setting considered in this Chapter, $D = (\mathbf{v}^s, \mathbf{v}^b)$ contains the reported valuations of sellers and buyers in the market. The algorithms we consider in this Chapter have output that can naturally be partitioned across the n users who provide the inputs — namely for each agent, whether they get to participate in a trade, and at what price. Let $\mathcal{M}$ be an algorithm that takes the data set D as input and outputs $\mathcal{M}(D) \in \mathcal{R}^n$, which is a vector whose ith coordinate corresponds to the output sent to agent i. Here $\mathcal{R}$ is the output range of the algorithm for a single agent, which we will take to be $\{0,1\} \times P$ (whether someone is chosen to participate in a trade, and a price for the trade). Informally speaking, differential privacy requires that a change in a single data entry should have little (distributional) effect on the output

[5] i.e., we only allocate agents willing to trade at price p. Agents not willing to participate at price p will opt out from the trade, thus are not taken into account by our benchmark.

[6] $\mathbf{1}[A]$, here and throughout the paper, represents the indicator function of event A.

of the mechanism. In other words, for every pair of data sets $D, D' \in \mathcal{D}^n$ that differ in at most one entry, differential privacy requires that the distribution of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ are "close" to each other where closeness is measured by the privacy parameters ε and δ.

Definition 3.2.2. Let $D, D' \in \mathcal{D}^n$ be two data sets of size n. We say D and D' are neighboring and write $D \sim D'$ if they differ in at most one data entry. D and D' are called i-neighbors ($D \sim_i D'$) if $D_{-i} = D'_{-i}$.

Definition 3.2.3 ((Standard) Differential Privacy (DP) [45])**.** An algorithm $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}^n$ is (ε, δ)-differentially private if for every pair of neighboring data sets $D \sim D' \in \mathcal{D}^n$, and for every subset of outputs $S \subseteq \mathcal{R}^n$,

$$\Pr\left[\mathcal{M}(D) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\mathcal{M}(D') \in S\right] + \delta$$

where the probability is taken with respect to the randomness of $\mathcal{M}$. if $\delta = 0$, $\mathcal{M}$ is said to be ε-DP.

We now define *joint differential privacy*. Joint differential privacy is defined in settings in which not only the inputs but also the outputs of the mechanism can be partitioned amongst the n users of the mechanism. In our setting, as in many mechanism design settings, this is the case: users report their valuations (which constitute the data) and then each receives an individual allocation. Joint differential privacy requires that an individual's input to the mechanism has little (distributional) effect on the outputs given to *others* — but allows one's own input to have a large effect on one's own output. Informally, it protects the privacy of each individual from arbitrary coalitions of other individuals using the system.

Definition 3.2.4 (Joint Differential Privacy [83])**.** An algorithm $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}^n$ is (ε, δ)-joint differentially private if for every i, for every pair of i-neighbors $D \sim_i D' \in \mathcal{D}^n$, and for every $S \subseteq \mathcal{R}^{n-1}$,

$$\Pr\left[\mathcal{M}(D)_{-i} \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[\mathcal{M}(D')_{-i} \in S\right] + \delta$$

where the probability is taken with respect to $\mathcal{M}$'s randomness. If $\delta = 0$, $\mathcal{M}$ is said to be ε-joint DP.

We will use the *Laplace* and *exponential mechanisms* of differential privacy in our proposed algorithms. See Appendix 3.7 for their formal definitions, their privacy and accuracy guarantees, and a few properties of differential privacy including *post-processing* and *composition*.

3.3 Private Call Auction Mechanisms

In this section, we outline our jointly differentially private mechanisms for the call auction problem and analyze their performance guarantees. Each mechanism's performance is measured in terms of its *payoff* — that is, the total number of shares cleared — as well as its *inventory* — the net position that the mechanism must itself take on. We measure our mechanisms' payoffs against the *maximal* number of shares that could be cleared with a uniform price, *given* the agents' reports.

Throughout this section, we assume reports are truthful; we will show in Section 3.4 that our mechanisms are approximately dominant strategy truthful. We also highlight that taking on *some* inventory is unavoidable – if the mechanism took no net position, a coalition of agents could use the constraint that the number of buyers and sellers must be equal to circumvent joint differential privacy – but our guarantees ensure that this net position remains small with high probability.

We propose three mechanisms. The first mechanism, described in Subsection 3.3.1, uses the exponential mechanism (see Appendix 3.7) to select a clearing price and then uses binomial randomization to determine who participates in a trade. In Subsection 3.3.2, we provide a second mechanism that again uses the exponential mechanism to select a price, but uses lottery numbers which are assigned to agents ex-ante to determine market participants. In Subsection 3.3.3, we describe a meta-algorithm that privately picks the mechanism with the better performance

guarantee,[7] and achieves performance as good as that of the best of the first two mechanisms.

Finally, in Subsection 3.3.4, we show matching lower bounds (up to log factors) for the payoff of *any* (ε, δ)-joint differentially private mechanism for the call auction.

3.3.1 A Private Call Auction Mechanism via Coin Flipping

In this subsection, we introduce our first jointly differentially private algorithm for selecting a price and allocating buyers and sellers to trades. The algorithm uses the exponential mechanism to differentially privately select a clearing price. With a slight abuse of notation, let

$$\Pi\left(p, \mathbf{v}^s, \mathbf{v}^b\right) = \min\left\{\sum_{i\in\mathcal{S}} \mathbf{1}\left[\mathbf{v}_i^s \leq p\right], \sum_{j\in\mathcal{B}} \mathbf{1}\left[\mathbf{v}_j^b \geq p\right]\right\} \tag{3.3.1}$$

be the number of trades that can happen at price p while guaranteeing every agent non-negative utility; $\Pi\left(p, \mathbf{v}^s, \mathbf{v}^b\right)$ is the utility function used by the exponential mechanism. After choosing the price, the mechanism randomly selects buyers and sellers willing to transact at the chosen price by flipping a coin with some particular bias for every agent in the market. The exchange then transacts with all selected transactors, possibly taking a net position in the asset. We formalize this mechanism in Algorithm 4. The mechanism takes data set $(\mathbf{v}^s, \mathbf{v}^b)$, privacy parameter ε, and confidence parameter α as inputs and outputs a price p and allocation vectors $\mathbf{a} = (\mathbf{a}^s, \mathbf{a}^b)$. In the algorithm description, $\exp(\cdot)$ is the exponential function, $Lap(\sigma)$ represents a mean-zero Laplace random variable with scale parameter σ, $(x)_+ := \max(x, 0)$, and $Bern(q)$ represents a Bernoulli random variable with success probability q.

We start the analysis by providing the privacy guarantees obtained by Algorithm 4:

[7] Which guarantee is best depends on the specific instance at hand.

ALGORITHM 4: Private Call Auction with Allocation via Coin Flipping ($\mathcal{M}_1$)

Input: Agents' valuations $(\mathbf{v}^s, \mathbf{v}^b)$, privacy level ε, confidence level α.

Output: Market price p, allocations $\mathbf{a} = (\mathbf{a}^s, \mathbf{a}^b)$.

Draw $p \propto \exp\left(\frac{\varepsilon \Pi(p, \mathbf{v}^s, \mathbf{v}^b)}{2}\right)$ ▷ Exponential mechanism chooses a price p privately

$\widehat{s} \leftarrow \sum_{i \in \mathcal{S}} \mathbf{1}\left[p \geq \mathbf{v}_i^s\right] + Lap(\frac{1}{\varepsilon})$ ▷ Privately estimate # of sellers willing to trade at p

$\widehat{b} \leftarrow \sum_{j \in \mathcal{B}} \mathbf{1}\left[p \leq \mathbf{v}_j^b\right] + Lap(\frac{1}{\varepsilon})$ ▷ Privately estimate # of buyers willing to trade at p

$\mathbf{a}_i^s \leftarrow \mathbf{1}\left[p \geq \mathbf{v}_i^s\right] \cdot Bern\left(q^s = \min\left\{1, \frac{\left(\widehat{b}\right)_+}{\left(\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon}\right)_+}\right\}\right)$ for all $i \in \mathcal{S}$. ▷ Sellers' allocations

$\mathbf{a}_j^b \leftarrow \mathbf{1}\left[p \leq \mathbf{v}_j^b\right] \cdot Bern\left(q^b = \min\left\{1, \frac{\left(\widehat{s}\right)_+}{\left(\widehat{b} - \frac{\ln(1/\alpha)}{\varepsilon}\right)_+}\right\}\right)$ for all $j \in \mathcal{B}$. ▷ Buyers' allocations

Claim 5. *The allocation mechanism described in Algorithm 4 satisfies 3ε joint differential privacy.*

The full proof of this claim can be found in Appendix 3.8. We also provide bounds on the payoff and the inventory of Mechanism 4 below:

Theorem 3.3.1 (Payoff and Inventory of Mechanism 4). Suppose $\text{opt} \geq 5\ln(V/\alpha)/\varepsilon$.

1. Payoff: with probability $1 - 8\alpha$,

$$\Pi\left(\mathcal{M}_1\right) \geq \text{opt} - \frac{2\ln(V/\alpha)}{\varepsilon} - \frac{2\ln\left(1/\alpha\right)}{\varepsilon} - \sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln\left(1/\alpha\right)}$$

2. Inventory: with probability $1 - 6\alpha$,

$$I\left(\mathcal{M}_1\right) \leq \frac{18\ln(1/\alpha)}{\varepsilon} + 2\sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(2/\alpha)} + \frac{4\ln(2/\alpha)}{3}$$

Remark 1. *Note that we constraint $\text{opt} = \Omega\left(\ln\left(V/\alpha\right)/\varepsilon\right)$. When $\text{opt} = O\left(\ln\left(V/\alpha\right)/\varepsilon\right)$, the inaccuracy introduced by releasing a differentially private price via the exponential mechanism is on the order of $\text{opt} = \Omega\left(\ln\left(V/\alpha\right)/\varepsilon\right)$, and we cannot hope to recover non-trivial utility guarantees.*

The proof of Theorem 3.3.1 is given in Appendix 3.9.1. We note that our bound does not follow directly from the classical guarantees of the Laplace and exponential mechanisms; it requires a more involved analysis of the concentration of the distribution of buyers and sellers selected to trade in Algorithm 4.

3.3.2 A Private Call Auction Mechanism via Lottery Numbers

Here, we present a second mechanism that, rather than using independent randomization to decide who participates in a trade, uses correlated randomization to improve the payoff and reduce the inventory requirements of the mechanism. In the second mechanism, participants are given data-independent "lottery numbers", and thresholds on these lottery numbers (selected using the exponential mechanism) are used to select among willing traders on both sides of the market. This correlation allows us to remove the $\sqrt{\text{opt}}$ term in the bounds of the previous mechanism, at the cost of introducing a logarithmic dependence on the number of agents n.

For a given valuation profile $(\mathbf{v}^s, \mathbf{v}^b)$, let $\Pi\left(p, \mathbf{v}^s, \mathbf{v}^b\right)$ be defined as in Equation 3.3.1. Assume seller i is assigned a lottery number $l_i^s \in [n^s]$ and buyer j is given $l_j^b \in [n^b]$ where we require that these lottery numbers are different for different agents. Without loss of generality, we assume $l_i^s = i$ and $l_j^b = j$. For a given price p, and profiles $(\mathbf{v}^s, \mathbf{v}^b)$, the loss of thresholds τ^s and τ^b on lottery numbers (one for sellers and one for buyers) is expressed as follows:

$$L^s\left(\tau^s, p, \mathbf{v}^s, \mathbf{v}^b\right) = \left|\sum_{i \in \mathcal{S}} \mathbf{1}\left[p \geq \mathbf{v}_i^s, \tau^s \geq i\right] - \Pi\left(p, \mathbf{v}^s, \mathbf{v}^b\right)\right|,$$

$$L^b\left(\tau^b, p, \mathbf{v}^b, \mathbf{v}^b\right) = \left|\sum_{j \in \mathcal{B}} \mathbf{1}\left[p \leq \mathbf{v}_j^b, \tau^b \leq j\right] - \Pi\left(p, \mathbf{v}^s, \mathbf{v}^b\right)\right|$$

For a price p, these loss functions measure how far off the number of agents chosen to trade on each side of the market would be from our target number of trades, $\Pi(p, \mathbf{v}^s, \mathbf{v}^b)$, if we used thresholds τ^s and τ^b as a tie-breaking rule to select sellers

and buyers who are willing to trade at price p, respectively. In Algorithm 5, just as before, we first use the exponential mechanism to select a price and then use the exponential mechanism with loss functions L^s and L^b (or utility functions: $-L^s$ and $-L^b$, based on the terminology used to describe the exponential mechanism in Appendix 3.7) to select the thresholds on lottery numbers.

ALGORITHM 5: Private Call Auction with Allocation via Lottery Numbers ($\mathcal{M}_2$)

Input: Agents' valuations $(\mathbf{v}^s, \mathbf{v}^b)$, privacy level ε.

Output: Market price p, allocations $\mathbf{a} = (\mathbf{a}^s, \mathbf{a}^b)$.

Draw $p \propto \exp\left(\frac{\varepsilon \Pi(p, \mathbf{v}^s, \mathbf{v}^b)}{2}\right)$ ▷ Exponential mechanism to privately choose a price p

Draw $\tau^s \propto \exp\left(-\frac{\varepsilon L^s(\tau^s, p, \mathbf{v}^s, \mathbf{v}^b)}{4}\right)$ ▷ Exponential mechanism to privately choose τ^s

Draw $\tau^b \propto \exp\left(-\frac{\varepsilon L^b(\tau^b, p, \mathbf{v}^s, \mathbf{v}^b)}{4}\right)$ ▷ Exponential mechanism to privately choose τ^b

$\mathbf{a}_i^s \leftarrow \mathbf{1}\left[p \geq \mathbf{v}_i^s, \tau^s \geq i\right]$ for all $i \in \mathcal{S}$. ▷ Sellers' allocations

$\mathbf{a}_j^b \leftarrow \mathbf{1}\left[p \leq \mathbf{v}_j^b, \tau^b \leq j\right]$ for all $j \in \mathcal{B}$. ▷ Buyers' allocations

Claim 6. *The allocation mechanism described in Algorithm 5 satisfies 3ε joint differential privacy.*

Theorem 3.3.2 (Payoff and Inventory of Mechanism 5). For any $\alpha > 0$,

1. Payoff: with probability $1 - 3\alpha$,

$$\Pi\left(\mathcal{M}_2\right) \geq \text{opt} - \frac{2\ln\left(V/\alpha\right)}{\varepsilon} - \frac{4\ln\left(n/\alpha\right)}{\varepsilon}$$

2. Inventory: with probability $1 - 2\alpha$,

$$I\left(\mathcal{M}_2\right) \leq \frac{8\ln\left(n/\alpha\right)}{\varepsilon},$$

The proof of Claim 6 is provided in Appendix 3.8, and that of Theorem 3.3.2 in Appendix 3.9.2.

3.3.3 A Meta Algorithm: Selecting the Best Mechanism Privately

Notice that the first term in the payoff bounds of both Theorems 3.3.1 and 3.3.2 are identical (as they both correspond to choosing a price using the exponential mechanism) but the remaining terms differ ($\mathcal{M}_1$ relies on binomial coin flips for tie-breaking whereas $\mathcal{M}_2$ tie-breaks via thresholds on lottery numbers). These two bounds are in general not comparable, as one depends on the maximum number of shares opt that can be cleared, whereas the other one depends on the total number n of agents in the market. The first bound provides better guarantees (up to constants and $\ln(1/\alpha)$ terms) when $\sqrt{\text{opt}} < \ln(n)/\varepsilon$, i.e. when the number of possible trades is significantly smaller than the total number of agents in the market,[8] whereas the second bound provides better guarantees when $\sqrt{\text{opt}} > \ln(n)/\varepsilon$.

We can achieve the better of the two bounds by comparing the bounds of Theorems 3.3.1 and 3.3.2 in a differentially-private manner and then running the mechanism with the better bound according to this private computation. To do so we compute the difference of payoff bounds of Mechanisms 4 and 5

$$f \triangleq \frac{2\ln(1/\alpha)}{\varepsilon} + \sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)} - \frac{4\ln(n/\alpha)}{\varepsilon}$$

in a differentially privately manner.[9] Then, based on the sign of f, the mechanism decides whether to run $\mathcal{M}_1$ or $\mathcal{M}_2$. The private computation of f will unavoidably add an extra term of order $\mathcal{O}(1/\varepsilon)$ to the final payoff bound. This mechanism is described in Algorithm 6. We provide guarantees on privacy, payoff, and inventory of this mechanism below.

[8] This models practical situations in repeated financial markets where sellers price a security higher than most buyers are willing to pay. In such situations, buyers may elect to wait until a new seller comes and offers a better price, while sellers may wait for a new buyer willing to buy at the current price.

[9] opt is a function of the input data set, hence a direct comparison of the bounds without addition of noise may leak information about the reported bids.

ALGORITHM 6: Private Call Auction with Allocation: A Meta Algorithm ($\mathcal{M}_3$)

Input: Agents' valuations $(\mathbf{v}^s, \mathbf{v}^b)$, privacy level ε, confidence level α.

Output: Market price p, allocations $\mathbf{a} = (\mathbf{a}^s, \mathbf{a}^b)$.

$f \leftarrow \frac{2\ln(1/\alpha)}{\varepsilon} + \sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)} - \frac{4\ln(n/\alpha)}{\varepsilon}$

$\widetilde{f} \leftarrow f + Lap\left(\frac{\sqrt{6\ln(1/\alpha)}}{\varepsilon}\right)$ ▷ Private estimate of f

if $\widetilde{f} < 0$ **then**

Run $\mathcal{M}_1\left(\mathbf{v}^s, \mathbf{v}^b, \varepsilon, \alpha\right)$ (Algorithm 4) and get $p, \mathbf{a}$.

else

Run $\mathcal{M}_2\left(\mathbf{v}^s, \mathbf{v}^b, \varepsilon\right)$ (Algorithm 5) and get $p, \mathbf{a}$.

end

Claim 7. *The allocation mechanism described in Algorithm 6 satisfies 7ε joint differential privacy.*

The proof of Claim 7 is provided in Appendix 3.8.

Theorem 3.3.3 (Payoff and Inventory of Mechanism 6). Suppose $\text{opt} \geq 5\ln(V/\alpha)/\varepsilon$.

1. Payoff: with probability $1 - 18\alpha$,

$$\Pi\left(\mathcal{M}_3\right) \geq \text{opt} - \frac{2\ln(\frac{V}{\alpha})}{\varepsilon} - \min\left\{\frac{2\ln\left(\frac{1}{\alpha}\right)}{\varepsilon} + \sqrt{6\left(\text{opt} + \frac{\ln(\frac{1}{\alpha})}{\varepsilon}\right)\ln\left(\frac{1}{\alpha}\right)}, \frac{4\ln(\frac{n}{\alpha})}{\varepsilon}\right\} - \frac{\sqrt{6}\ln^{1.5}(\frac{1}{\alpha})}{\varepsilon}$$

2. Inventory: with probability $1 - 14\alpha$,

$$I\left(\mathcal{M}_3\right) \leq 4\min\left\{\frac{2\ln\left(\frac{1}{\alpha}\right)}{\varepsilon} + \sqrt{6\left(\text{opt} + \frac{\ln\left(\frac{1}{\alpha}\right)}{\varepsilon}\right)\ln\left(\frac{1}{\alpha}\right)}, \frac{4\ln(\frac{n}{\alpha})}{\varepsilon}\right\} + \frac{4\sqrt{6}\ln^{1.5}(\frac{1}{\alpha})}{\varepsilon} + \frac{10\ln(\frac{1}{\alpha})}{\varepsilon} + \frac{4\ln(\frac{2}{\alpha})}{3}$$

This theorem follows from Theorems 3.3.1 and 3.3.2, as well as the accuracy guarantee of the Laplace mechanism used in Algorithm 6 to compute f. We defer the full proof to Appendix 3.9.3.

3.3.4 A Lower Bound

We now provide a lower bound showing that *any* algorithm which computes a price in an (ε, δ)-differentially private manner and allocates among willing participants at this price *must*, for *some* instance, suffer a loss of $\Omega\left(1/\varepsilon\right)$ (compared to the optimal number of shares that could be cleared on that instance). Because this bound applies to a broader set of mechanisms that reveal *only* the price privately (but may select the optimal allocation absent privacy), it also applies to the mechanisms considered in Section 3.3. We will compare the performance of any given differentially private algorithm on several input data sets. To do so, we will define an instance-dependent benchmark below, that we call $\text{opt}(D)$. Formally, given an input data set $D = (\mathbf{v}^s, \mathbf{v}^b)$, our benchmark is:

$$\text{opt}\left(D\right) = \max_{p} \min\left\{\sum_{i \in \mathcal{S}} \mathbf{1}\left[\mathbf{v}_i^s \leq p\right], \sum_{j \in S} \mathbf{1}\left[\mathbf{v}_j^b \geq p\right]\right\}.$$

Definition 3.3.1 (Loss of an algorithm)**.** For any (possibly randomized) algorithm $\mathcal{A} : \mathcal{D}^n \to P$ that takes a data set $D = (\mathbf{v}^s, \mathbf{v}^b)$ as an input and outputs a price p, the loss of $\mathcal{A}$ on input data set $D = (\mathbf{v}^s, \mathbf{v}^b)$ of agents valuations is defined as follows:

$$L\left(\mathcal{A}, D\right) = \text{opt}\left(D\right) - \text{E}_{p \sim \mathcal{A}(D)}\left[\min\left\{\sum_{i \in \mathcal{S}} \mathbf{1}\left[\mathbf{v}_i^s \leq p\right], \sum_{j \in S} \mathbf{1}\left[\mathbf{v}_j^b \geq p\right]\right\}\right].$$

I.e., this loss compares the number of trades that could be cleared in expectation at the price selected by $\mathcal{A}$ to the maximum number of trades when the trading price is optimally chosen. We define the worst-case expected loss of $\mathcal{A}$ as the worst-case loss over all data sets, i.e. $L\left(\mathcal{A}\right) = \sup_D\left[L(\mathcal{A}, D)\right]$.

Our lower bound will hold so long as δ is not too large in comparison with ε.[10] We note that our lower bound on the expected loss matches the $\widetilde{\mathcal{O}}\left(1/\varepsilon\right)$ dependencies[11]

[10] Typically, differentially private algorithms use $\delta << \varepsilon$.

[11] The instances we construct use $V, n \sim 1/\varepsilon$. The logarithmic dependencies of our upper bounds in n and V translate into logarithmic dependencies in $1/\varepsilon$, hence the $\widetilde{\mathcal{O}}$ notation.

of our high probability upper bounds on the loss for Mechanisms 4, 5, 6 (and consequently of any upper bound on the expected loss of these mechanisms). Finally, it is worth remarking that our lower bound for (ε, δ)-DP mechanisms matches the upper bound obtained by restricting attention to $(\varepsilon, 0)$-DP mechanisms; this implies that relaxing δ-privacy requirements of Mechanisms 4, 5, 6 will not lead to any significant improvements in terms of their accuracy guarantees.

Theorem 3.3.4. [Lower bound on the loss of private algorithms] Pick any ε, δ such that $0 \leq \varepsilon \leq 1$ and $\delta = \mathcal{O}(\varepsilon)$. There exists a range of (integer) valuations $P(\varepsilon)$ and a number of agents $n(\varepsilon)$ such that *any* (ε, δ)-DP algorithm $\mathcal{A} : \mathcal{D}^{n(\varepsilon)} \to P(\varepsilon)$ must suffer worst-case expected loss of $\Omega\left(1/\varepsilon\right)$.

The proof of Theorem 3.3.4 relies on constructing a family of data sets $\{D_l\}_l$ such that no differentially private algorithm $\mathcal{A}$ can simultaneously suffer expected loss of $\mathcal{O}\left(1/\varepsilon\right)$ on all of them. We do so by carefully calibrating the following trade-off: on the one hand, we require any pair of data sets in $\{D_l\}_l$ be close enough that the stability properties of differential privacy guarantee any private algorithm must pick a similar distribution of prices on both data sets. On the other hand, we require that the data sets furthest from each other are different *enough* such that no fixed distribution can incur a low loss on both.

3.3.5 Connections to the Market Impact Literature

As mentioned in the Introduction, it is possible to draw some informal but interesting connections between this work and the finance literature on market impact. Market impact models typically propose strong stochastic assumptions on price formation (e.g. random walk and diffusion models or martingale assumptions on limit order dynamics) and then solve for the optimal strategy to minimize trading costs and price impact. In particular, there is a large body of work on the so-called "square root law" (see, eg. [61, 62]). which predicts that the change to price inflicted by a trade of k shares scales with $\sqrt{k/\mathcal{V}}$, where $\mathcal{V}$ is the total volume of

shares cleared during the trade; the ratio $k/\mathcal{V}$ is referred to as the trade's *participation rate*. As we note below, $\mathcal{V}$ is typically closely related to other measures of market activity such as the number of orders placed (as with our n) or the number of quote changes in limit order dynamics.

Our results imply that the change in the expected clearing price in our DP call auction resulting from an order of k shares is bounded by a multiplicative factor of $(e^{k\varepsilon} - 1)$. Setting this equal to $\sqrt{k/n}$ to match the square root law[12] and solving for ε approximately yields $\varepsilon \approx 1/\sqrt{kn}$ for small participation rates. Plugging this into our utility bound of Theorem 3.3.2, the shares we execute at this ε scales like $\text{opt}(1-\sqrt{kn}/\text{opt})$. Thus as long as k is $o(n)$ and opt scales with n,[13] asymptotically we approach opt with the same price impact as that predicted by the square root law but with two major advantages. First, we have made *no* assumptions, stochastic or otherwise, on the orders placed by market participants. Second, we are not only bounding the price impact, we are also bounding information leakage of *any* form, as per the promises of differential privacy.

3.4 Strategic Framework

In Section 3.3, we focused on the algorithmic form of our mechanism and provided privacy guarantees and optimality guarantees with respect to the reported valuations, without regard to whether those reports are truthful or not. In this section, we embed our mechanisms into a game theoretic framework and examine its properties, including (approximate) truthfulness. More precisely, we now assume the agents are strategic; they may decide to report a bid that differs from their valu-

[12] Here we are assuming that the number of orders n in our model plays the role of $\mathcal{V}$ above; see subsequent footnote.

[13] This scaling is broadly consistent with recent data from electronic exchanges. For instance, the ratio of shares traded to quote changes (a common measure of market activity) across 3443 U.S. equities averaged 0.16 with standard deviation 0.09.

ation, or even to not participate in the mechanism in the first place. Formally, all sellers i and buyers j have quasi-linear utilities determined by their own valuations and the outcome of the mechanism: $\mathbf{u}_i^s(\mathcal{M}) = \mathbf{a}_i^s \cdot (p - \mathbf{v}_i^s), \mathbf{u}_j^b(\mathcal{M}) = \mathbf{a}_i^b \cdot (\mathbf{v}_j^b - p)$, where, with a slight abuse of notation, we omit the dependency of $\mathcal{M}$ on the agents' reports.

Buyers and sellers aim to maximize their utility from participating (or not participating) in the mechanism. In the face of strategic behavior, we will require our mechanisms to be (approximately) truthful and individually rational; i.e., it should never be in an agent's best interest to misreport his valuation, and an agent should always have a strategy that guarantees non-negative utility from participating in the mechanism and so would rather participate than not. Individual rationality and (approximate) truthfulness are formally defined below:

Definition 3.4.1 (Ex-Post Individual Rationality)**.** We say a double-auction mechanism $\mathcal{M}$ satisfies ex-post individual rationality if, for every seller $i \in \mathcal{S}$, there exists a bid $\mathbf{r}_i^s$ for agent i such that for every possible set of bids $\mathbf{r}_{-i}$ submitted by all agents but i, and every realization of the randomness of the mechanism $\mathcal{M}$, $\mathbf{u}_i^s\left(\mathcal{M}\left(\mathbf{r}_i^s, \mathbf{r}_{-i}\right)\right) \geq 0$, and similarly for every buyer j, there exists a bid $\mathbf{r}_j^b$ for agent j such that for every possible set of bids $\mathbf{r}_{-j}$ submitted by all agents but j, and every realization of the randomness of the mechanism $\mathcal{M}$, $\mathbf{u}_j^b\left(\mathcal{M}\left(\mathbf{r}_j^b, \mathbf{r}_{-j}\right)\right) \geq 0$.

Definition 3.4.2 (Approximate Dominant-Strategy Truthfulness)**.** We say a double-auction mechanism $\mathcal{M}$ satisfies γ-approximate dominant-strategy truthfulness if, for every seller $i \in \mathcal{S}$, every possible bid $\mathbf{r}_i^s$ submitted by i, and every possible set of bids $\mathbf{r}_{-i}$ submitted by all agents but i,

$$\mathrm{E}_{\mathcal{M}}\left[\mathbf{u}_i^s\left(\mathcal{M}\left(\mathbf{r}_i^s, \mathbf{r}_{-i}\right)\right)\right] \leq \mathrm{E}_{\mathcal{M}}\left[\mathbf{u}_i^s\left(\mathcal{M}\left(\mathbf{v}_i^s, \mathbf{r}_{-i}\right)\right)\right] + \gamma$$

and similarly for every buyer j, for every possible bid $\mathbf{r}_j^b$ submitted by j and every possible set of bids $\mathbf{r}_{-j}$ submitted by all agents but j,

$$\mathrm{E}_{\mathcal{M}}\left[\mathbf{u}_j^b\left(\mathcal{M}\left(\mathbf{r}_j^b, \mathbf{r}_{-j}\right)\right)\right] \leq \mathrm{E}_{\mathcal{M}}\left[\mathbf{u}_j^b\left(\mathcal{M}\left(\mathbf{v}_j^b, \mathbf{r}_{-j}\right)\right)\right] + \gamma$$

where expectations are taken with respect to the randomness of $\mathcal{M}$.

In Section 3.4.1, we show that our mechanisms are *individual rational* and (unlike in the standard call auction) approximately *dominant-strategy truthful*. While our results assume that agents wish to trade a single share, we show how our per-share guarantees translate into (gracefully degrading) per-player guarantees in more general setting in which agents can trade multiple shares.

Then, in Sections 3.4.2-3.4.2, we consider *learning dynamics* under both the standard call auction and our mechanism and show that a system in which agents use a modified exponential weights algorithm (which we call "Social" Exponential Weights) to learn to bid will eventually converge to the optimal number of shares cleared. While it is true that truthfulness implies that agents cannot do better than bidding their true values, one might consider learning dynamics for two reasons. First, if agents do not trust the mechanism designer (or share their assumptions), applying a no-regret learning algorithm is a plausible response to guarantee good performance. Second, good outcomes obtained in the presence of decentralized, distributed, and selfish algorithms are compelling evidence of the robustness and quality of our mechanism. To our knowledge, the use of no-regret learning algorithms by all agents in a call auction setting has not been studied before, and these results may be of independent interest.

3.4.1 Individual Rationality and Truthfulness Properties of Our Algorithms

In this section, we discuss the incentive properties of our proposed algorithms. To do so, we note that $(\mathbf{v}^s, \mathbf{v}^b)$ are the true valuations of sellers and buyers and denote their revealed bids by $(\mathbf{r}^s, \mathbf{r}^b)$, respectively. To study truthfulness, we assume seller i (buyer j) can submit a bid $\mathbf{r}_i^s$ ($\mathbf{r}_j^b$) that may not be equal to their valuation $\mathbf{v}_i^s$ ($\mathbf{v}_j^b$), and show that it is approximately never in agent i's (resp. j's) best interest

to do so. We start by noting that our mechanisms are individually rational:

Claim 8 (Individual Rationality)**.** *The mechanisms described in Algorithms 4, 5, and 6 are ex-post individually rational.*

Proof. We prove the result for Mechanism 4; proofs for the other mechanisms are similar. It suffices to show that there exists a strategy for any seller (resp. any buyer) that guarantees him non-negative utility. For any seller i, setting $\mathbf{r}_i^s = \mathbf{v}_i^s$ is a strategy that ensures that whenever i is allocated a trade (i.e. $\mathbf{a}_i^s = 1$), it must be that $p \geq \mathbf{r}_i^s = \mathbf{v}_i^s$; this immediately guarantees i gets non-negative utility—independently of how other agents bid and of the randomness of the mechanism. A similar proof holds for buyers. □

We also show that differential privacy guarantees approximate truthfulness in the dominant-strategy sense: i.e., it does not allow agents (sellers and buyers) to gain too much profit by submitting a bid different than their true valuation, *no matter what the realized bids of the other agents are*[14].

Claim 9 (Approximate Truthfulness)**.** *The mechanisms described in Algorithms 4 and 5 satisfy γ-approximate dominant-strategy truthfulness for $\gamma = (e^{3\varepsilon} - 1)V$; the mechanism described in Algorithm 6 satisfies γ-approximate dominant-strategy truthfulness for $\gamma = (e^{7\varepsilon} - 1)V$.*

We defer the full proof to Appendix 3.11. In the proof, we first observe that since the market price is chosen subject to differential privacy, individual agents cannot significantly change it by misreporting their valuations. However, this is not enough to argue truthfulness, as under *joint* differential privacy, an agent's allocation may heavily depend on his report. To complete the proof, we show that the function by

[14]Truthfulness is desirable not only because it makes computing equilibrium strategies and predicting equilibrium behavior simpler, but also because knowing the *true* valuations allows the mechanism designer to clear the most shares.

which the mechanism determines transactors is a best-response for an agent with the reported valuation given the output of the differentially private mechanism.

Note that, in general, call auctions are *not* dominant-strategy truthful, since even small bidders may impact the price selected by a mechanism acting on reported bids. This is a consequence of the fact that in the simple call auction (as well as in continuous order book mechanisms) the optimal price is, in general, *not* stable [53]. Importantly, we note that the truthfulness guarantees are a function of ε; as opt grows larger, ε can be made smaller with less and less relative cost. Consequently, the truthfulness guarantee can be made stronger for a given level of privacy as the number of optimal trades cleared increases.

We highlight that because our strategic framework assumes each bidder controls a single share, our guarantees are at the *per-share* level. Our privacy guarantees generalize, however, to the case where bidders control at most k shares by expanding ε by a factor of k.[15] Our truthfulness guarantees also follow by expanding ε by a factor of k.

3.4.2 Learning in Repeated Call Auctions

In this section, we consider a *repeated* call auction. Agents are initially unaware of each other's valuations and behavior and run simple learning algorithms to learn how to bid. In each time step t, each seller i (respectively buyer j) reports a bid $\mathbf{r}^s_{i,t}$ (resp. $\mathbf{r}^b_{j,t}$), which may differ from his valuation, to the mechanism. Given this input $(\mathbf{r}^s_t, \mathbf{r}^b_t)$, the mechanism computes and publicly releases a price p_t and assigns an allocation $\mathbf{a}_{i,t}$ to each seller i (respectively $\mathbf{a}_{j,t}$ to each buyer j). We will consider two versions of this mechanism, one that is non-private and is inspired by standard

[15] An ε-differentially private mechanism with respect to a single share is $k\varepsilon$-differentially private with respect to the data of a bidder who controls k shares; intuitively, this is because an agent that misreports his valuation over k shares creates a dataset that is a k-neighbor of the dataset in which they had bid truthfully. Such a bidder can affect the distribution of prices by an amount of at most $e^{k\varepsilon}$, and so their own expected utility by $(e^{k\varepsilon} - 1)V$.

call auctions, in Section 3.4.2, and one that is private and is based on Mechanism 4, in Section 3.4.2. The agents then update their bidding strategies based on the quantities outputted by the mechanism, via a simple no-regret algorithm (Exponential Weights).

We highlight that our agents are *naive* in that they do not compute a counterfactual price p_t and allocation vector $\mathbf{a}_t$ given alternative bids they could have made. Instead, they only update their bidding strategies with respect to how much better off they could have been by bidding differently, *assuming they had no effect on the price*. The motivation for this is two-fold: first, counterfactual reasoning would require the agents to know the bids of other agents, which are not released by the mechanism (and typically not available in many real-life call auctions). Second, when agents are small relative to the total market, they may believe that their actions *do not* greatly affect these quantities. We note that differential privacy makes this belief into a *property* of our mechanism rather than a naive assumption. Thus, small bidders using naive updates will have a *real* regret guarantee when interacting with a differentially private call auction.

Learning in the Absence of Privacy

In this section, we focus on learning dynamics when the mechanism runs a standard call auction, absent privacy; this non-private setting will serve as a natural point of comparison for dynamics with respect to our private mechanism. At every time step t, agents submit bids that may differ from their valuations. In response, the mechanism computes a price and allocation, with the goal of maximizing traded shares among willing participants. We denote the agents' reports as $\mathbf{r}^s_{i,t}$ and $\mathbf{r}^b_{j,t}$ for seller i and buyer j, respectively, at time t. The mechanism chooses a price p_t to maximize

$$\Pi\left(p, \mathbf{r}^s_t, \mathbf{r}^b_t\right) \triangleq \min\left\{\sum_{i\in\mathcal{S}} \mathbf{1}\left[\mathbf{r}^s_{i,t} \leq p\right], \sum_{j\in\mathcal{B}} \mathbf{1}\left[\mathbf{r}^b_{j,t} \geq p\right]\right\},$$

which is the number of shares the mechanism will trade at price p, assuming that sellers will only agree to trade when the price is above their reported bid and buyers when the price is below their reported bid. To compute the allocation $\mathbf{a}_t = (\mathbf{a}_t^s, \mathbf{a}_t^b)$, the mechanism must choose among these sellers and buyers it believes (based on the reports) are willing to trade at the chosen price p_t. When there are an equal number of sellers and buyers willing to trade at price p_t, the mechanism allocates a trade to all of them; otherwise, the mechanism randomly selects a subset of $\Pi(p, \mathbf{r}_t^s, \mathbf{r}_t^b)$ agents from the side with excess number of willing participants. Formally, the mechanism computes $q_t^b = \Pi(p_t, \mathbf{r}_t^s, \mathbf{r}_t^b) / \sum_{j \in \mathcal{B}} \mathbf{1}[\mathbf{r}_{j,t}^b \geq p_t]$ and $q_t^s = \Pi(p_t, \mathbf{r}_t^s, \mathbf{r}_t^b) / \sum_{i \in \mathcal{S}} \mathbf{1}[\mathbf{r}_{i,t}^s \leq p_t]$; these probabilities will be less than 1 on the excess side of the market and exactly 1 on the short side. We assume the mechanism publicly releases p_t, q_t^s, and q_t^b to all agents in the market, and communicates to each seller i (resp. buyer j) his own allocation $\mathbf{a}_{i,t}^s$ (resp. $\mathbf{a}_{j,t}^b$).

Agents learn via Exponential Weights: A natural no-regret (*regret* here is the classic notion of performance in online learning) algorithm for updating bidding strategies is the Exponential Weights mechanism. We describe the classical *Exponential Weights Update* rule for buyers (buyer j) in Algorithm 7, and note that this update is defined symmetrically for the sellers.

ALGORITHM 7: Exponential Weights

Input: Learning rate η.

Set $\mathbf{w}_{j,1}^b(k) \leftarrow \frac{1}{\mathbf{v}_j^b}$ for $k = 1, \ldots, \mathbf{v}_j^b$ ▷ Initialize uniform weights.

for $t \in 1...T$ **do**

$\mathbf{r}_{j,t}^b \sim \mathbf{w}_{j,t}^b$ ▷ Draw bid from distribution

$\mu_{j,t}^b(k) \leftarrow q_t^b(\mathbf{v}_j^b - p_t)\mathbf{1}[k \geq p_t]$ for $k = 1, \ldots, \mathbf{v}_j^b$ ▷ Observe payoff of each bid k

$\mathbf{w}_{j,t+1}^b(k) \leftarrow \frac{\exp(\eta \mu_{j,t}^b(k))}{\sum_j \mathbf{w}_{j,t}^b(j) \exp(\eta \mu_{j,t}^b(j))} \cdot \mathbf{w}_{j,t}^b(k)$ for $k = 1, \ldots, \mathbf{v}_j^b$ ▷ Update the weights.

end

Informally, the updates work as follows. Initially, we assume every seller bids

uniformly above their value and every buyer bids uniformly below their value.[16] Then, in each round t, for every possible $k \in P$, agents compute what their expected payoff would have been had they reported k as their valuation, given the *current* price p_t and the allocation probabilities q_t^s and q_t^b. They use these expected payoffs to update their distribution of bids, in a way that puts exponentially more weight on bids with higher expected utilities; the speed at which these updates happen is controlled by the learning rate parameter η, taken here to be constant. For appropriate choices of learning rate η, this algorithm is known to be no-regret.

One may hope these dynamics converge to clearing opt shares with probability going to 1, where opt is defined as in Equation 3.2.1. However, this may not be the case when agents update their weights according to Algorithm 7. This stems from the fact that agents are indifferent between trading at their valuation, and not trading at all, as both net a payoff of zero. This is reflected in the exponential weight update, and buyers learn to put a significant amount of weight on bids that are strictly less than their valuation (as trades for those bids are strictly profitable). When clearing opt trades requires many agents to bid exactly at their valuation, the number of shares cleared is bounded away from the benchmark. We show instead that the dynamics will clear the following benchmark, which only considers trades that are strictly profitable for both sides of the market:

Definition 3.4.3 (Optimal Jointly Profitable Trades)**.** We let opt$'$ be the maximum number of trades achievable for a given $(\mathbf{v}^s, \mathbf{v}^b)$, such that all trading buyers and sellers get strictly positive utility. Formally,

$$\text{opt}' = \max_p \min \left\{ \sum_{i \in \mathcal{S}} \mathbf{1}\left[\mathbf{v}_i^s < p\right], \sum_{j \in \mathcal{B}} \mathbf{1}\left[\mathbf{v}_j^b > p\right] \right\}.$$

[16] A buyer j cannot improve his utility by bidding over his valuation (as increasing his bid cannot decrease p_t nor increase his probability of allocation), and risks obtaining negative utility by doing so, if $\mathbf{v}_{j,t}^b < p_t \leq \mathbf{r}_{j,t}^b$. Hence, bidding above his valuation is a dominated strategy for the buyer. Similarly, bidding under his value is a dominated strategy for a seller. The assumption of this prior knowledge can be relaxed at the price of slower convergence.

We call this benchmark the "Optimal Jointly Profitable Trades with Uniform Pricing" benchmark.

The statement showing that the mechanism will converge in probability to clearing at least opt′ shares is formalized below:

Theorem 3.4.1 (Convergence to (at least) opt′). Suppose buyers and sellers update their bid distributions according to Algorithm 7 (with any $\eta > 0$). Further, at any time t, suppose p_t is chosen uniformly at random among the set of optimal prices at time t. Then, the number of shares cleared at time t satisfies

$$\lim_{t\to\infty} \Pr\left[\Pi\left(p_t, \mathbf{r}_t^s, \mathbf{r}_t^b\right) \geq \text{opt}'\right] = 1.$$

We also provide a variant of the Exponential Weights algorithm, that we will show converges to opt shares cleared. This variant is described in Algorithm 8.

ALGORITHM 8: Social Exponential Weights

Input: Learning rate η, "fake" utility ξ.

$\mathbf{w}_{j,1}^b(k) \leftarrow 1/\mathbf{v}_j^b$ for $k = 1, \ldots, \mathbf{v}_j^b$ ▷ Initialize with uniform weights.

for $t = 1, \ldots, T$ **do**

  $\mathbf{r}_{j,t}^b \sim \mathbf{w}_{j,t}^b$ ▷ Draw from the current weights.

  if $\mathbf{v}_j^b \neq p_t$ **then**

    $\mu_{j,t}^b(k) \leftarrow q_t^b(\mathbf{v}_j^b - p_t)\mathbf{1}[k \geq p_t]$ for $k = 1, \ldots, \mathbf{v}_j^b$ ▷ expected utility for bid k.

  else

    $\mu_{j,t}^b(k) \leftarrow q_t^b \xi \mathbf{1}[k = \mathbf{v}_j^b]$ for $k = 1, \ldots, \mathbf{v}_j^b$ ▷ Agent pretends getting utility ξ from trading.

  end

  $\mathbf{w}_{j,t+1}^b(k) \leftarrow \frac{\exp(\eta\mu_{j,t}^b(k))}{\sum_l \mathbf{w}_{j,t}^b(l)\exp(\eta\mu_{j,t}^b(l))} \cdot \mathbf{w}_{j,t}^b(k)$ for $k = 1, \ldots, \mathbf{v}_j^b$ ▷ Update the weights.

end

Algorithm 8 is a modification of the classic Exponential Weights algorithm. In particular, when the price is equal to agent's valuation, the algorithm assigns a nonzero utility $q_t^b\xi$ to reporting the agent's valuation, for ξ arbitrarily small; this can

be seen as agents updating their weights as if they strictly preferred trading to not trading, even when their trade would make no profit. In other words, it implements a preference to break ties (in utility) in favor of trading over not trading. We call this "Social" Exponential Weights because incorporating this modified utility allows the system as a whole to reach a better social outcome (one with more shares traded) than otherwise. Crucially, despite this modification, Algorithm 8 remains no-regret for a fixed horizon T with appropriate choices of learning rate, η, and "fake" utility, ξ:

Lemma 3.4.4 (No-regret)**.** *Algorithm 8 is no-regret ($\mathcal{O}\left(\sqrt{T}\right)$ cumulative regret) for $\eta, \xi = \mathcal{O}\left(1/\sqrt{T}\right)$.*

The proof is almost identical to that of the no-regret guarantees of traditional exponential weights, and is deferred to Appendix 3.12.1. We highlight that we define regret with respect to the single best action in hindsight given the *fixed* sequence of prices observed; that is, we do not consider the notion of *Stackelberg* regret, which is calculated with respect to the best fixed action *given* that the mechanism picks a sequence of prices in response to the selected actions (see, e.g. [43]). If agents are small enough that their actions do not greatly affect the mechanism's responses, then the standard notion of regret and Stackelberg regret do not greatly differ; if, moreover, a mechanism is differentially private, then (for small enough agents) these notions of regret coincide, because differential privacy ensures that agents placing small orders have little impact on the price.

Under Algorithm 8, the number of shares cleared converges to opt with probability that tends to 1 as t grows large. We make this statement formally below:

Theorem 3.4.2 (Convergence to opt). Suppose buyers and sellers that update their bidding strategies according to Algorithm 8 (with any $\eta, \xi > 0$). Further, suppose p_t is chosen uniformly at random among the set of optimal prices at time t. Then,

the number of shares cleared at time t satisfies

$$\lim_{t\to\infty} \Pr\left[\Pi\left(p_t, \mathbf{r}_t^s, \mathbf{r}_t^b\right) = \text{opt}\right] = 1.$$

To prove this result, we show that with a small, constant probability (in t) in any given round, all agents bid their valuations. In such cases, the mechanism picks an optimal price, and at least OPT buyers (resp. sellers) increase their probability of bidding above (resp. below) this price. When the number of rounds goes to infinity, this event is repeated infinitely often for some optimal price $p^\star$, and OPT buyers (resp. sellers) bid above (resp. below) $p^\star$ with probability that tends to 1. The full proof is given in Appendix 3.12.3. A similar argument is used to prove Theorem 3.4.1, in Appendix 3.12.2.

Learning in Repeated Call Auctions with Differential Privacy

We now consider the same dynamic setting as before, with the difference that the centralized designer now computes the price p_t and the allocation $\mathbf{a}_t$ at time t in a joint-differentially private fashion. For simplicity of exposition, we pick the private mechanism used by the designer to be Mechanism 4, which picks a price via the exponential mechanism and picks agents to allocate from the smaller side of the market via binomial coin flips. We show that when agents play according to the exponential weights (resp. Social EW) algorithm, the dynamics converge to clearing at least opt (resp. opt$'$) shares minus inaccuracies introduced by privacy.

Theorem 3.4.3. Suppose buyers and sellers update their bidding strategies according to Algorithm 8 (with any $\eta, \xi > 0$). Further, suppose the market allocation mechanism is Algorithm 4. There exists an integer $N(\alpha)$ such that for any $t \geq N(\alpha)$, the number of shares cleared at time t satisfies

$$\Pr\left[\Pi\left(p_t, \mathbf{r}_t^s, \mathbf{r}_t^b\right) \geq \text{opt} - \frac{2\ln(V/\alpha)}{\varepsilon} - \frac{2\ln(1/\alpha)}{\varepsilon} - \sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)}\right] \geq 1 - 9\alpha.$$

where this probability is taken with respect to the randomness of both Algorithms 4 and 8.

The proof idea is the following: despite the price randomness due to privacy, the event in which all agents bid their value and an optimal price is picked happens infinitely often, as in the non-private case. In turn, (at least) OPT buyers (resp. sellers) eventually learn to bid above (resp. below) an optimal price $p^\star$. However, the mechanism will still pick sub-optimal prices to guarantee privacy, as per the bound of Theorem 3.3.1. We refer the reader to Appendix 3.12.4 for a complete proof. A similar statement holds, with respect to benchmark opt$'$ (see Definition 3.4.3), when agents update according to Algorithm 7.

Theorem 3.4.4. Suppose buyers and sellers use the Exponential Weights Algorithm 7 (with any $\eta > 0$) to update their bids. Further, suppose the market allocation mechanism is Algorithm 4. There exists an integer $N(\alpha) > 0$ such that for all $t \geq N(\alpha)$, the number of shares cleared at time t satisfies

$$\Pr\left[\Pi\left(p_t, \mathbf{r}_t^s, \mathbf{r}_t^b\right) \geq \text{opt}' - \frac{2\ln(V/\alpha)}{\varepsilon} - \frac{2\ln\left(1/\alpha\right)}{\varepsilon} - \sqrt{6\left(\text{opt}' + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln\left(1/\alpha\right)}\right] \geq 1 - 9\alpha.$$

3.5 Simulations

In previous sections, we designed our mechanism and obtained theoretical guarantees of performance; these guarantees were given both in a *one-shot* setting and relative to the optimal result that could be reached given agents' bids and also in a *repeated* setting using no-regret learning. In this section, we conduct experiments on simulated data in both a one-shot and learning setting in order to explore how tightly these guarantees bind in practice.

We perform all simulations in MATLAB using a similar starting configuration. We have 5000 buyers and 5000 sellers, and valuations must be integer values between 1 and 100. Valuations are drawn from normal distributions centered at 45 for sellers and 55 for buyers, with standard deviations of 15 for both. The draws are rounded to the nearest integer, and draws below 1 or above 100 are replaced with 1 and 100 respectively.

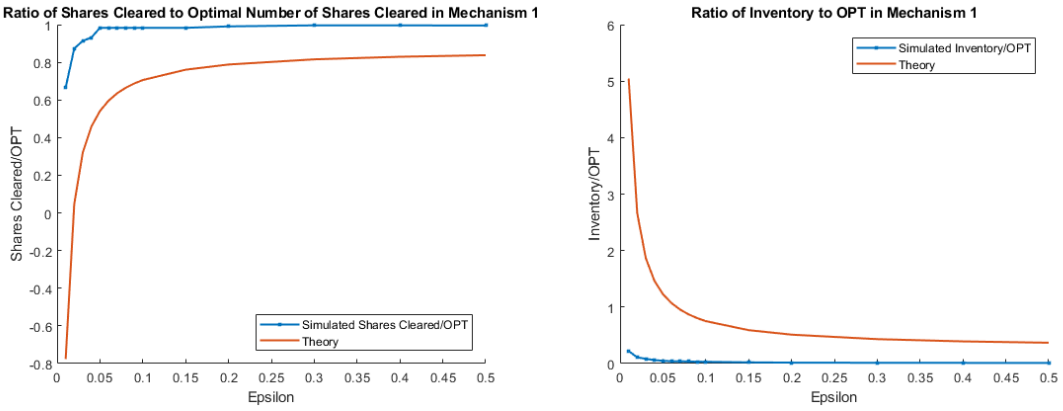


Figure 3.1: Realized payoff and inventory relative to theoretical optimal in one-shot game for varying ε.

The mechanism we implement is the first we defined (Algorithm 4), run once or repeatedly for the one-shot game and learning settings, respectively. We vary ε over a range from $\varepsilon = 0.01$ to $\varepsilon = 0.5$.

Single-shot game For the single shot game, we perform 800 trials per value of ε with a fixed set of agent valuations. These valuations were drawn randomly according to the procedure described above. We assume agents bid truthfully (and all of our comparisons are to the truthful optimal).

In the first plot of Figure 3.1, we show the empirical 5% quantile (i.e. the value for which only 5% of draws saw lower values) of the *competitive ratio* defined as the shares cleared as a fraction of opt, the optimal number of shares that can be cleared given the realized valuations. This competitive ratio quantile is plotted in blue. The appropriate guarantee to compare to is the lower bound on this quantile given in Theorem 3.3.1, with confidence parameter $\alpha = 0.05/8$; this bound is plotted in orange. While the realized competitive ratio indicates, unsurprisingly, that privacy is not costless for very small levels of ε, it remains far above the worst-case guarantee predicted, and rapidly increases to nearly 1 in the practical regime (i.e., even for $\varepsilon = 0.1$). This shows that, for a large enough number of agents and valuations drawn

from well-behaved distributions, reasonable privacy can be achieved in practice with very little loss in utility.

The second plot of Figure 3.1 shows the inventory taken on by the mechanism, again plotting this quantile as a ratio of the optimal number of shares cleared in blue (again, limited to the top 95% of runs) and the theoretical upper bound for $\alpha = 0.05/6$ (as per the inventory bound of Theorem 3.3.1) in orange. Notice that for very small ε, the theoretical guarantee can be extremely large; yet, again, the realized inventory is far below the guarantee and never exceeds 23% for even $\varepsilon = 0.01$ and is less than 5% for $\varepsilon \geq 0.05$.

Learning Setting In the *learning* setting, we plot the shares cleared *over time* as agents learn to bid given their valuations. We repeat the auction for 1000 rounds (1500 for the imbalance plot) with learning rate $\eta = 0.1$ and "fake" utility $\xi = 0.1$. Agents draw fixed valuations and then use the Social Exponential Weights described in Algorithm 8 to learn and bid each round. We repeat this process for several different values of ε.

The first three plots in Figure 3.2 tell similar stories: agents, and thus the system, learn to bid over time in such a way as to clear the optimal number of shares (were the mechanism privacy-free). The noisiness in the plots is due to privacy and depends on the choice of ε: the smaller the value of ε, the more likely the mechanism is to pick a sub-optimal price, even after agents learn to bid optimally. For $\varepsilon = 0.01$, the randomness of the mechanism induces enough noise as to occasionally forego a large portion of utility; at larger values of ε, the added randomness costs relatively little.

The fourth plot displays the imbalance between number of buyers bidding above vs. sellers bidding below the price chosen by the repeated *standard* (i.e., non-private) call auction when buyers use Social Exponential Weights. We highlight an interesting connection to real-world behavior: NYSE and NASDAQ perform pre-

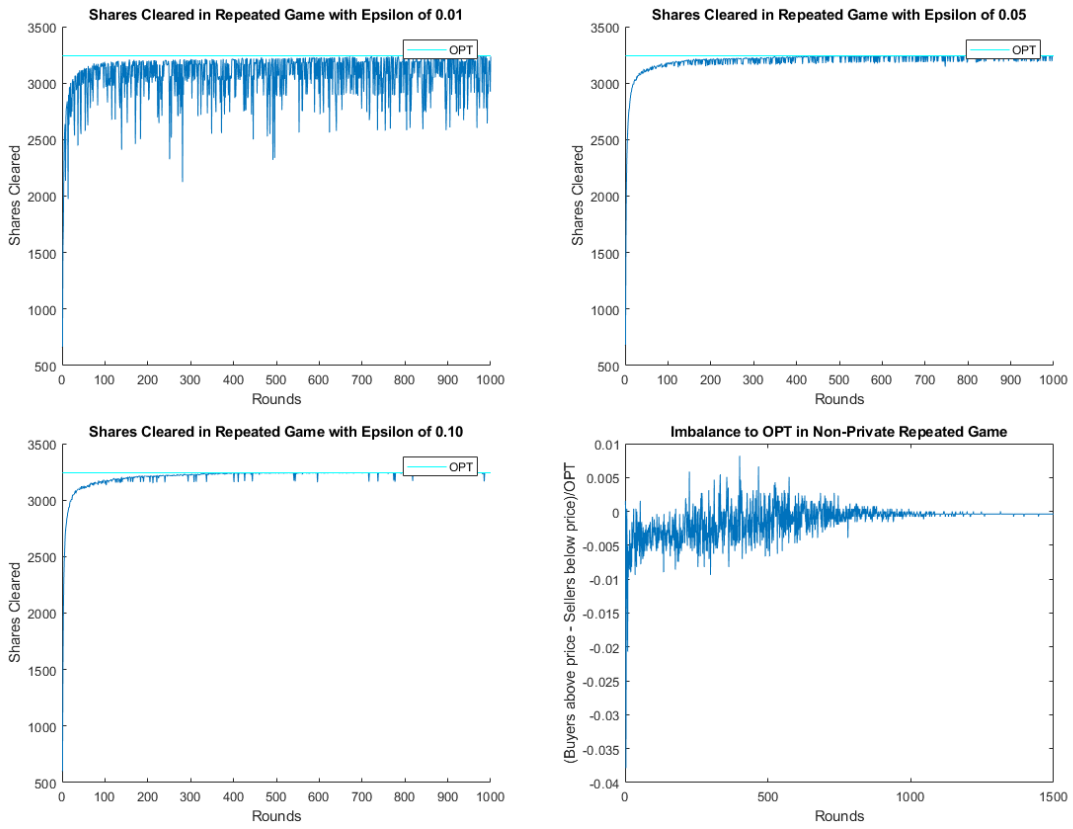


Figure 3.2: The first three plots show the shares cleared over time, in the repeated setting of our mechanism, using Social Exponential Weights (8) for various choices of ε. The last plot shows the imbalance between buyers and sellers over time in a repeated (non-private) auction. The agents' updates use $\eta, \xi = 0.1$.

opening or pre-closing repeated "hypothetical" auctions aimed at price discovery. In these hypothetical auctions, the exchanges accept bids, announce the current price and imbalance, allow bidders to submit updated bids, and repeat. The pattern in imbalances documented by [30] agrees broadly with that of Figure 3.2: that is, the imbalance begins skewed to one side or another, but it repeatedly oscillates as bidders adjust before converging to a settled state.

3.6 Appendix to Chapter 3

3.7 Differential Privacy Tools

In this section, we remind the reader of mechanisms that are classically used to guarantee differential privacy. These mechanisms work by adding appropriately-chosen noise to the choices and outputs of a mechanism, so as to ensure that a change in a single individual's data cannot have a large distributional effect on the mechanism's output. The noise introduced by differentially private mechanisms depends not only on the level (ε, δ) of privacy one aims to guarantee, but also on the *sensitivity* of the query of interest. This sensitivity measures how much the real-valued function of interest is affected by a change in a single entry of an input data set, and will be formally defined in our introduced DP mechanisms.

A commonly used mechanism for releasing the answer to numerical queries while guaranteeing $(\varepsilon, 0)$-differential privacy is the Laplace mechanism. The Laplace mechanism takes a numerical query f as an input, and perturbs the value of f on the input data set with zero-mean Laplace noise that has scale proportional to $(\Delta f/\varepsilon)$ where Δf is the ℓ_1-sensitivity of f.

Definition 3.7.1 (Laplace Mechanism [45])**.** Given a function $f : \mathcal{D}^n \to \mathbb{R}^k$ with ℓ_1-sensitivity Δf:

$$\Delta f = \max_{\substack{D, D' \in \mathcal{D}^n \\ D \sim D'}} \|f(D) - f(D')\|_1 ,$$

a data set $D \in \mathcal{D}^n$, and a privacy parameter ε, the Laplace mechanism outputs:

$$f_\varepsilon (D) = f (D) + (W_1, \ldots, W_k)$$

where W_i's are *i.i.d.* random variables drawn from $\text{Lap}\,(\Delta f/\varepsilon)$.

We provide the privacy and accuracy guarantees of the Laplace mechanism below:

Theorem 3.7.1 (Privacy vs. Accuracy of the Laplace Mechanism [45]). The Laplace Mechanism guarantees $(\varepsilon, 0)$-differential privacy and that with probability at least $1-\delta$,

$$\|f_\varepsilon(D) - f(D)\|_\infty \leq \ln\left(\frac{k}{\delta}\right) \cdot \left(\frac{\Delta f}{\varepsilon}\right)$$

We remark that the Laplace mechanism can be used to privately output the answer to numerical queries. However, suppose we want to privately output the solution to a maximization problem defined on the input data. Then, directly adding noise to the optimal solution could completely destroy the objective value of the maximization problem in question (for example, in an auction, adding a small amount of noise on the price of an item could significantly reduce revenue). In such situations, the Laplace mechanism performs poorly, and a better choice of private mechanism is the Exponential Mechanism, defined below:

Definition 3.7.2 (Exponential Mechanism [94])**.** Let $U : \mathcal{D}^n \times P \to \mathbb{R}$ be a utility function that takes a data set $D \in \mathcal{D}^n$ and a parameter $p \in P$ as inputs, and let ΔU be its sensitivity. In other words,

$$\Delta U = \max_{p \in P} \max_{\substack{D, D' \in \mathcal{D}^n \\ D \sim D'}} |U(D, p) - U(D', p)|.$$

Given a data set $D \in \mathcal{D}^n$ and a privacy parameter ε, the exponential mechanism outputs $p \in P$ with probability proportional to $\exp\left(\frac{\varepsilon U(D,p)}{2\Delta U}\right)$ where $\exp(\cdot)$ is the exponential function.

Theorem 3.7.2 (Privacy vs. Accuracy of the Exponential Mechanism [94]). The Exponential Mechanism guarantees $(\varepsilon, 0)$-differential privacy. Further, let $p_\varepsilon \in P$ be the output of the Exponential mechanism, we have that with probability at least $1-\delta$,

$$\left|U(D, p_\varepsilon) - \max_{p \in P} U(D, p)\right| \leq \ln\left(\frac{|P|}{\delta}\right) \cdot \left(\frac{2\Delta U}{\varepsilon}\right)$$

An important property of differential privacy is that it is robust to *post-processing*. Applying any data-independent function to the output of an (ε, δ)-DP algorithm preserves (ε, δ)-differential privacy.

Lemma 3.7.3 (Post-Processing [45]). *Let $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}$ be an (ε, δ)-DP algorithm and let $f : \mathcal{R} \to \mathcal{R}'$ be any function. We have that the algorithm $f \, o \, \mathcal{M} : \mathcal{D}^n \to \mathcal{R}'$ is (ε, δ)-DP.*

Another important property of differential privacy is that DP algorithms can be composed adaptively with a graceful degradation in their privacy parameters.

Theorem 3.7.3 ((Simple) Composition [47]). Let $\mathcal{M}_t$ be an $(\varepsilon_t, \delta_t)$-DP algorithm for $t \in [T]$. We have that the composition $\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_T)$ is (ε, δ)-DP where $\varepsilon = \sum_t \varepsilon_t$ and $\delta = \sum_t \delta_t$.

To prove that our mechanisms satisfy (ε, δ)-joint differential privacy, we will leverage the billboard lemma ([72]). The billboard lemma shows that for every individual i in the data set, restricting i's output to be a function of only the output of a differentially private mechanism (run on all agents' data) and his own input guarantees joint differential privacy.

Lemma 3.7.4 (Billboard Lemma [72]). *Suppose $\mathcal{M} : \mathcal{D}^n \to \mathcal{R}'$ is (ε, δ)-differentially private. Consider any set of functions $f_i : \mathcal{D}_i \times \mathcal{R}' \to \mathcal{R}$, where $\mathcal{D}_i$ is the portion of the data set containing i's data. The composition $\{f_i(\Pi_i(D), \mathcal{M}(D))\}$ is (ε, δ)-jointly differentially private, where $\Pi_i : \mathcal{D}^n \to \mathcal{D}_i$ is the projection to i's data.*

3.8 Proofs of Privacy guarantees of our mechanisms

proof of Claim 5. We start the proof by noticing that the sensitivity of Π (as per Definition 3.7.2) is 1: indeed, changing one element in the data $(\mathbf{v}^s, \mathbf{v}^b)$, i.e. the valuation of a single agent, will change the number of shares cleared by at most one, for any fixed price p. We can therefore conclude that by Theorem 3.7.2 the mechanism that takes the data set as input and outputs a price $p \propto \exp\left(\varepsilon \Pi(p, \mathbf{v}^s, \mathbf{v}^b)/2\right)$ is ε-DP. One can similarly argue that given a fixed price p, quantities $\sum_{i \in \mathcal{S}} \mathbf{1}[p \geq \mathbf{v}_i^s]$

and $\sum_{j\in\mathcal{B}} \mathbf{1}[p \le \mathbf{v}_j^b]$ have sensitivity 1 (see Definition 3.7.1), and therefore by Theorem 3.7.1, $\widehat{s}$ and $\widehat{b}$ both satisfy ε-DP. We can now invoke the Composition Theorem 3.7.3 to conclude that the triplet $(p, \widehat{s}, \widehat{b})$ computed in Algorithm 4 satisfies 3ε-differential privacy. The claim then follows by the Billboard Lemma 3.7.4 and noticing that each agent's allocation depends only on their own data and the triplet $(p, \widehat{s}, \widehat{b})$. □

Proof of Claim 6. Notice first that according to Definition 3.7.2, the sensitivity of Π is 1 because for any p, changing one agent's valuation can change Π by at most 1. Now fixing the price p output by the first exponential mechanism, it similarly follows that the sensitivity of loss functions L^s and L^b are 2. We therefore have that the exponential mechanisms outputting p, τ^s, and τ^b are all ε-DP by Theorem 3.7.2, and hence the triplet (p, τ^s, τ^b) satisfies 3ε-DP by the Composition Theorem 3.7.3. The claim then follows by the Billboard Lemma 3.7.4 and noticing that each agent's allocation depends only on their own data and the triplet (p, τ^s, τ^b). □

Proof of Claim 7. First, we note that the sensitivity of f is upper-bounded by $\sqrt{6\ln(1/\alpha)}$. Indeed, we remind the reader that opt has sensitivity of 1, and note that for any $x \ge 0$

$$\sqrt{6\ln(1/\alpha)} \ge \sqrt{6\left(x + 1 + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)} - \sqrt{6\left(x + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)},$$

using the classical inequality $\sqrt{a} + \sqrt{b} \ge \sqrt{a+b}$. Therefore, by Theorem 3.7.1, the computation of $\widetilde{f}$ is ε-differential private. In each of mechanisms $\mathcal{M}_1$ and $\mathcal{M}_2$, p_1 the price output by $\mathcal{M}_1$, respectively p_2 the price output by $\mathcal{M}_2$, are computed in an ε-differentially private manner. Similarly, $\hat{s}$, $\hat{b}$ (resp. τ^s, τ^b), the private counts of the number of agents willing to trade in $\mathcal{M}_1$ at price p_1 (resp. the thresholds picked by mechanism $\mathcal{M}_2$ for price p_2) are each the result of an ε-differentially private conputation (conditional on p_1, p_2). In turn, our mechanism can be seen as one that computes $(\widetilde{f}, p_1, p_2, \hat{s}, \hat{b}, \tau^s, \tau^b)$ in a 7ε-differentially private manner (by

the composition guarantee of Theorem 3.7.3), then outputs an allocation $\mathbf{a}_i^s$ for each given seller i (resp. $\mathbf{a}_j^b$ for each buyer j) as a function of only $\mathbf{v}_i^s$ (resp. $\mathbf{v}_j^b$) and $(\widetilde{f}, p_1, p_2, \hat{s}, \hat{b}, \tau^s, \tau^b)$. Hence, by Lemma 3.7.4, $\mathcal{M}$ is 7ε-joint differentially private. □

3.9 Proofs of Profit and Inventory of our Mechanisms

3.9.1 Proof of Theorem 3.3.1

Proof. We will be using the following concentration inequalities in our proof.

Fact 3.9.1 (Multiplicative Chernoff Bound)**.** Let $\{X_i\}_{i=1}^n$ be a collection of independent random variables where $X_i \in [0,1]$ for all i. Let $S = \sum_{i=1}^n X_i$ and $\mu = \mathrm{E}[S]$. We have that for any $0 \leq t \leq 1$,

$$\Pr\left[S < (1-t)\mu\right] \leq e^{-\frac{\mu t^2}{2}}$$

Fact 3.9.2 (Bernstein's Inequality)**.** Let $\{X_i\}_{i=1}^n$ be a collection of $i.i.d$ random variables where for each i, $X_i \in [0,1]$, $E[X_i] = \mu$, and $Var(X_i) = \sigma^2$. Let $S = \sum_{i=1}^n X_i$. We have that for any $t \geq 0$,

$$\Pr\left[|S - n\mu| > t\right] \leq 2e^{-\frac{t^2}{2n\sigma^2 + 2t/3}}$$

Let $s(p) = \sum_{i \in S} \mathbf{1}\left[p \geq \mathbf{v}_i^s\right]$ and $b(p) = \sum_{j \in B} \mathbf{1}\left[p \leq \mathbf{v}_j^b\right]$ be the number of sellers and buyers available at price p, where p is the price chosen by the exponential mechanism. Note that as p is a random variable, so are $s(p)$ and $b(p)$. From now on, for simplicity of notations, we omit the dependency of s and b in p. We start the proof by noting that by the accuracy guarantee of the Laplace mechanism (see Theorem 3.7.1) and a union bound, we have with probability at least $1 - 2\alpha$ that

$$\left|\widehat{b} - b\right| \leq \frac{\ln(1/\alpha)}{\varepsilon}, \quad \left|\widehat{s} - s\right| \leq \frac{\ln(1/\alpha)}{\varepsilon}, \tag{3.9.1}$$

and by the accuracy guarantee of the Exponential mechanism (see Theorem 3.7.2), we have with probability at least $1-\alpha$ that,

$$\Pi\left(p, \mathbf{v}^s, \mathbf{v}^b\right) = \min\left\{s, b\right\} \geq \text{opt} - \frac{2\ln(V/\alpha)}{\varepsilon}. \tag{3.9.2}$$

By a union bound, Equations (3.9.1) and (3.9.2) hold simultaneously with probability at least $1-3\alpha$, and throughout this proof we condition on these events. Let $\widetilde{s} = \sum_{i\in\mathcal{S}} \mathbf{a}_i^s$ and $\widetilde{b} = \sum_{j\in\mathcal{B}} \mathbf{a}_j^b$ be the number of sellers and buyers who participate in a trade, output by the mechanism. First, let's focus on $\widetilde{s}$. Observe that

$$\widetilde{s}\,|s, \widehat{s}, \widehat{b} \sim \mathit{Binomial}\left(s, \widehat{q} = \min\left\{1, \frac{\left(\widehat{b}\right)_+}{\left(\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon}\right)_+}\right\}\right).$$

Note that we have

$$\begin{aligned}
\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon} &\geq s - \frac{2\ln(1/\alpha)}{\varepsilon} && \text{(by Equation (3.9.1))}\\
&\geq \text{opt} - \frac{2\ln(1/\alpha)}{\varepsilon} - \frac{2\ln(V/\alpha)}{\varepsilon} && \text{(by Equation (3.9.2))}\\
&\geq \text{opt} - \frac{4\ln(V/\alpha)}{\varepsilon} && (V \geq 1)\\
&\geq \frac{\ln(V/\alpha)}{\varepsilon} && \text{(by assumption, } OPT \geq \tfrac{5\ln(V/\alpha)}{\varepsilon}\text{)}\\
&> 0,
\end{aligned}$$

and also

$$\begin{aligned}
\widehat{b} &\geq b - \frac{\ln(1/\alpha)}{\varepsilon} && \text{(by Equation (3.9.1))}\\
&\geq \text{opt} - \frac{\ln(1/\alpha)}{\varepsilon} - \frac{2\ln(V/\alpha)}{\varepsilon} && \text{(by Equation (3.9.2))}\\
&\geq \text{opt} - \frac{3\ln(V/\alpha)}{\varepsilon} && (V \geq 1)\\
&\geq \frac{2\ln(V/\alpha)}{\varepsilon} && \text{(by assumption, } OPT \geq \tfrac{5\ln(V/\alpha)}{\varepsilon}\text{)}\\
&> 0.
\end{aligned}$$

As such, $\widehat{q}$ is well-defined, and we can rewrite it as

$$\widehat{q} = \min\left(1, \frac{\widehat{b}}{\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon}}\right).$$

We have by the multiplicative Chernoff Bound (Fact 3.9.1) with $t = \sqrt{\frac{2\ln(1/\alpha)}{s\widehat{q}}}$ that,

$$\widetilde{s} \geq s\widehat{q} - \sqrt{2s\widehat{q}\ln(1/\alpha)} \tag{3.9.3}$$

with probability at least $1-\alpha$ when $t \leq 1$. Note that the bound applies when $t > 1$ too, noting that then $s\widehat{q} - \sqrt{2s\widehat{q}\ln(1/\alpha)} < 0$ but $\widetilde{s} \geq 0$. In what follows, we will provide an upper bound and a lower bound for the term $s\widehat{q}$ so that we can further lower bound $\widetilde{s}$ in Equation (3.9.3). Symmetrically, we can get a similar lower bound for $\widetilde{b}$ which completes the first part of the proof because $\Pi(\mathcal{M}) = \min\{\widetilde{s}, \widetilde{b}\}$.

On the one hand, note that

$$\begin{aligned}
s\widehat{q} &= s \cdot \frac{\min\left\{\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon}, \widehat{b}\right\}}{\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon}} \\
&\geq \frac{s}{s} \cdot \min\left\{\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon}, \widehat{b}\right\} \\
&= \min\left\{\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon}, \widehat{b}\right\} \\
&\geq \min\{s, b\} - \frac{2\ln(1/\alpha)}{\varepsilon} \\
&\geq \mathrm{opt} - \frac{2\ln(1/\alpha)}{\varepsilon} - \frac{2\ln(V/\alpha)}{\varepsilon}.
\end{aligned} \tag{3.9.4}$$

The first inequality follows from $\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon} \leq s$ by Equation (3.9.1). The second inequality follows from

$$\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon} \geq s - \frac{2\ln(1/\alpha)}{\varepsilon}, \quad \widehat{b} \geq b - \frac{\ln(1/\alpha)}{\varepsilon}$$

by Equation (3.9.1). The third inequality is a direct application of Equation (3.9.2).

On the other hand,

$$\begin{aligned} s\widehat{q} &= s \cdot \frac{\min\left\{\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon}, \widehat{b}\right\}}{\widehat{s} - \frac{\ln(1/\alpha)}{\varepsilon}} \\ &\leq \frac{s}{s - \frac{2\ln(1/\alpha)}{\varepsilon}} \cdot \min\left\{s, b + \frac{\ln(1/\alpha)}{\varepsilon}\right\} \\ &\leq \frac{s}{s - \frac{2\ln(1/\alpha)}{\varepsilon}} \cdot \left(\min\{s, b\} + \frac{\ln(1/\alpha)}{\varepsilon}\right) \\ &\leq \frac{s}{s - \frac{2\ln(V/\alpha)}{\varepsilon}} \cdot \left(\min\{s, b\} + \frac{\ln(1/\alpha)}{\varepsilon}\right) \\ &\leq 3\left(\min\{s, b\} + \frac{\ln(1/\alpha)}{\varepsilon}\right) \\ &\leq 3\left(\mathrm{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right). \end{aligned} \tag{3.9.5}$$

The first inequality follows from

$$s - \frac{\ln(1/\alpha)}{\varepsilon} \leq \widehat{s} \leq s + \frac{\ln(1/\alpha)}{\varepsilon}, \quad \widehat{b} \leq b + \frac{\ln(1/\alpha)}{\varepsilon}$$

by Equation (3.9.1). The second-to-last inequality follows from the fact that

$$f : (0, +\infty) \to \mathbb{R}, \quad f(x) = \frac{x}{x - \frac{2\ln(V/\alpha)}{\varepsilon}} = \frac{1}{1 - \frac{2\ln(V/\alpha)}{\varepsilon x}}$$

is a non-increasing function of x, and that by Equation (3.9.2),

$$s \geq \mathrm{opt} - \frac{2\ln(V/\alpha)}{\varepsilon} \geq \frac{3\ln(V/\alpha)}{\varepsilon}.$$

Combining Equations (3.9.3), (3.9.4), and (3.9.5), we obtain that with probability $1 - 4\alpha$,

$$\widetilde{s} \geq \mathrm{opt} - \frac{2\ln(1/\alpha)}{\varepsilon} - \frac{2\ln(V/\alpha)}{\varepsilon} - \sqrt{6\left(\mathrm{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)}. \tag{3.9.6}$$

Symmetrically, we can get the same bound for $\widetilde{b}$: with probability $1 - 4\alpha$,

$$\widetilde{b} \geq \mathrm{opt} - \frac{2\ln(1/\alpha)}{\varepsilon} - \frac{2\ln(V/\alpha)}{\varepsilon} - \sqrt{6\left(\mathrm{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)} \tag{3.9.7}$$

Combining Equations (3.9.6) and (3.9.7) and noting that $\Pi(\mathcal{M}_1) = \min\{\widetilde{s}, \widetilde{b}\}$ proves the first part of the theorem. We conclude the proof by noting that the statements hold with probability at least $1 - 8\alpha$ by union bound.

Let us now analyze the inventory of the mechanism. We have by the triangle inequality that

$$I\left(\mathcal{M}_1\right) = \left|\widetilde{s} - \widetilde{b}\right| \leq \left|\widetilde{s} - \min\left\{s, b\right\}\right| + \left|\widetilde{b} - \min\left\{s, b\right\}\right| \tag{3.9.8}$$

We will provide an upper bound for the first term, and by symmetry, an upper bound on the second will follow immediately. We have that by the triangle inequality

$$\left|\widetilde{s} - \min\left\{s, b\right\}\right| \leq \left|\widetilde{s} - s\widehat{q}\right| + \left|s\widehat{q} - \min\left\{s, b\right\}\right| \tag{3.9.9}$$

First

$$\begin{aligned}
\left|\widetilde{s} - s\widehat{q}\right| &\leq \sqrt{2s\widehat{q}(1-\widehat{q})\ln(2/\alpha)} + \frac{2\ln(2/\alpha)}{3} \\
&\leq \sqrt{2s\widehat{q}\ln(2/\alpha)} + \frac{2\ln(2/\alpha)}{3} \\
&\leq \sqrt{6\left(\mathrm{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(2/\alpha)} + \frac{2\ln(2/\alpha)}{3}
\end{aligned} \tag{3.9.10}$$

where the first inequality holds with probability $1 - \alpha$ and follows from the Bernstein's inequality (Fact 3.9.2) with $\sigma^2 = \widehat{q}(1-\widehat{q})$ and taking $t = \frac{\ln(2/\alpha)}{3} + \sqrt{\frac{\ln^2(2/\alpha)}{9} + 2n\sigma^2\ln(2/\alpha)}$. Notice that $t \leq \frac{2\ln(2/\alpha)}{3} + \sqrt{2n\sigma^2\ln(2/\alpha)}$. The last inequality follows from the upper bound developed in Equation (3.9.5). Second,

$$\left|s\widehat{q} - \min\left\{s, b\right\}\right| \leq \frac{9\ln(1/\alpha)}{\varepsilon}. \tag{3.9.11}$$

This is because as we showed in Equation (3.9.4),

$$-\frac{2\ln\left(1/\alpha\right)}{\varepsilon} \leq s\widehat{q} - \min\left\{s, b\right\},$$

and as we showed in Equation (3.9.5),

$$\begin{aligned}
s\widehat{q} - \min\{s,b\} &\leq \frac{s}{s - \frac{2\ln(1/\alpha)}{\varepsilon}} \cdot \left(\min\{s,b\} + \frac{\ln(1/\alpha)}{\varepsilon}\right) - \min\{s,b\} \\
&= \left(\frac{s}{s - \frac{2\ln(1/\alpha)}{\varepsilon}} - 1\right) \cdot \min\{s,b\} + \frac{s}{s - \frac{2\ln(1/\alpha)}{\varepsilon}} \cdot \frac{\ln(1/\alpha)}{\varepsilon} \\
&\leq \left(\frac{2\ln(1/\alpha)/\varepsilon}{s - \frac{2\ln(1/\alpha)}{\varepsilon}}\right) \cdot s + \frac{s}{s - \frac{2\ln(1/\alpha)}{\varepsilon}} \cdot \frac{\ln(1/\alpha)}{\varepsilon} \\
&= \frac{s}{s - \frac{2\ln(1/\alpha)}{\varepsilon}} \cdot \frac{3\ln(1/\alpha)}{\varepsilon} \\
&\leq \frac{9\ln(1/\alpha)}{\varepsilon}
\end{aligned}$$

Because we showed in Equation (3.9.5) that $\frac{s}{s - \frac{2\ln(1/\alpha)}{\varepsilon}} \leq 3$. Putting together Equations (3.9.9), (3.9.10), and (3.9.11) we get that with probability $1 - 3\alpha$,

$$|\widetilde{s} - \min\{s,b\}| \leq \sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(2/\alpha)} + \frac{2\ln(2/\alpha)}{3} + \frac{9\ln(1/\alpha)}{\varepsilon}.$$

Swapping the roles of the buyers and a similar proof yields the same bound on $\left|\widetilde{b} - \min\{s,b\}\right|$. A union bound completes the proof, by Equation (3.9.8). □

3.9.2 Proof of Theorem 3.3.2

Proof. Let us define $\widetilde{s} = \sum_{i \in \mathcal{S}} \mathbf{a}_i^s$ and $\widetilde{b} = \sum_{j \in \mathcal{B}} \mathbf{a}_j^b$ to be the number of sellers and buyers who participate in a trade under output allocation $\mathbf{a}$. We have that with probability at least $1-3\alpha$, by the accuracy guarantee of the Exponential mechanism (see Theorem 3.7.2),

$$\Pi\left(p, \mathbf{v}^s, \mathbf{v}^b\right) \geq \text{opt} - \frac{2\ln(V/\alpha)}{\varepsilon}, \tag{3.9.12}$$

and that since $\min_{\tau^s} L^s(\tau^s, p, \mathbf{v}^s, \mathbf{v}^b) = \min_{\tau_b} L^b(\tau^b, p, \mathbf{v}^s, \mathbf{v}^b) = 0$,

$$\left|\widetilde{s} - \Pi(p, \mathbf{v}^s, \mathbf{v}^b)\right| = L^s(\tau^s, p, \mathbf{v}^s, \mathbf{v}^b) \leq \frac{4\ln\left(n^s/\alpha\right)}{\varepsilon} \quad \Longrightarrow \quad \widetilde{s} \geq \Pi(p, \mathbf{v}^s, \mathbf{v}^b) - \frac{4\ln\left(n^s/\alpha\right)}{\varepsilon}, \tag{3.9.13}$$

$$\left|\widetilde{b} - \Pi(p, \mathbf{v}^s, \mathbf{v}^b)\right| = L^b(\tau^b, p, \mathbf{v}^s, \mathbf{v}^b) \leq \frac{4\ln\left(n^b/\alpha\right)}{\varepsilon} \quad \Longrightarrow \quad \widetilde{b} \geq \Pi(p, \mathbf{v}^s, \mathbf{v}^b) - \frac{4\ln\left(n^b/\alpha\right)}{\varepsilon} \tag{3.9.14}$$

We therefore have that

$$\Pi\left(\mathcal{M}_2\right) = \min\left\{\widetilde{s}, \widetilde{b}\right\} \geq \text{opt} - \frac{2\ln(V/\alpha)}{\varepsilon} - \frac{4\ln(n/\alpha)}{\varepsilon} \tag{3.9.15}$$

Let's now analyze the inventory introduced by the private mechanism. We have that

$$\begin{aligned} I\left(\mathcal{M}_2\right) &= \left|\widetilde{s} - \widetilde{b}\right| \\ &\leq \left|\widetilde{s} - \Pi\left(p, \mathbf{v}^s, \mathbf{v}^b\right)\right| + \left|\widetilde{b} - \Pi\left(p, \mathbf{v}^s, \mathbf{v}^b\right)\right| \\ &= L^s(\tau^s, p, \mathbf{v}^s, \mathbf{v}^b) + L^b(\tau^b, p, \mathbf{v}^s, \mathbf{v}^b) \\ &\leq \frac{4\ln\left(n^s/\alpha\right)}{\varepsilon} + \frac{4\ln\left(n^b/\alpha\right)}{\varepsilon} \\ &\leq \frac{8\ln(n/\alpha)}{\varepsilon} \end{aligned}$$

where the second inequality holds with probability $1 - 2\alpha$ by Equations (3.9.13) and (3.9.14). □

3.9.3 Proof of Theorem 3.3.3

Proof. This theorem follows from Theorems 3.3.1 and 3.3.2 and conditioning on the accuracy guarantee of the additional Laplace mechanism used in Algorithm 6:

$$\text{w.p. } 1 - \alpha, \qquad \left|\widetilde{f} - f\right| \leq \frac{\sqrt{6}\ln^{1.5}(1/\alpha)}{\varepsilon} \tag{3.9.16}$$

Suppose $\widetilde{f} < 0$. Note that in this case,

$$\begin{aligned}
\text{opt} - \Pi(\mathcal{M}_3) &= \text{opt} - \Pi(\mathcal{M}_1) \\
&\leq \frac{2\ln(V/\alpha)}{\varepsilon} + \frac{2\ln(1/\alpha)}{\varepsilon} + \sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)} \qquad (\star) \\
&= \frac{2\ln(V/\alpha)}{\varepsilon} + \frac{4\ln(n/\alpha)}{\varepsilon} + f \\
&\leq \frac{2\ln(V/\alpha)}{\varepsilon} + \frac{4\ln(n/\alpha)}{\varepsilon} + \widetilde{f} + \frac{\sqrt{6}\ln^{1.5}(1/\alpha)}{\varepsilon} \\
&\leq \frac{2\ln(V/\alpha)}{\varepsilon} + \frac{4\ln(n/\alpha)}{\varepsilon} + \frac{\sqrt{6}\ln^{1.5}(1/\alpha)}{\varepsilon} \qquad (\star\star)
\end{aligned}$$

where the first inequality follows from Theorem 3.3.1, with probability $1 - 8\alpha$. The second inequality follows from Equation 3.9.16, with probability $1 - \alpha$. Combining the bounds given by the second and the last inequalities (specified by $\star$ and $\star\star$), we get that with probability $1 - 9\alpha$,

$$\begin{aligned}
\text{opt} - \Pi(\mathcal{M}_3) \leq \min\left\{\frac{2\ln(1/\alpha)}{\varepsilon} + \sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)}, \frac{4\ln(n/\alpha)}{\varepsilon}\right\} \\
+ \frac{2\ln(V/\alpha)}{\varepsilon} + \frac{\sqrt{6}\ln^{1.5}(1/\alpha)}{\varepsilon}.
\end{aligned}$$

A similar analysis for $\widetilde{f} \geq 0$ which uses Theorem 3.3.2 gives us the same bound and proves the first part of the theorem.

Now let's look at the inventory. Suppose $\widetilde{f} \leq 0$. We have that

$$\begin{aligned}
I(\mathcal{M}_3) &= I(\mathcal{M}_1) \\
&\leq \frac{18\ln(1/\alpha)}{\varepsilon} + 2\sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(2/\alpha)} + \frac{4\ln(2/\alpha)}{3} \\
&\leq \frac{18\ln(1/\alpha)}{\varepsilon} + 4\sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)} + \frac{4\ln(2/\alpha)}{3} \qquad (\star) \\
&= 4\left(f + \frac{4\ln(n/\alpha)}{\varepsilon}\right) + \frac{10\ln(1/\alpha)}{\varepsilon} + \frac{4\ln(2/\alpha)}{3} \\
&\leq 4\left(\widetilde{f} + \frac{\sqrt{6}\ln^{1.5}(1/\alpha)}{\varepsilon} + \frac{4\ln(n/\alpha)}{\varepsilon}\right) + \frac{10\ln(1/\alpha)}{\varepsilon} + \frac{4\ln(2/\alpha)}{3} \\
&\leq 4\left(\frac{\sqrt{6}\ln^{1.5}(1/\alpha)}{\varepsilon} + \frac{4\ln(n/\alpha)}{\varepsilon}\right) + \frac{10\ln(1/\alpha)}{\varepsilon} + \frac{4\ln(2/\alpha)}{3} \qquad (\star\star)
\end{aligned}$$

where the first inequality follows from Theorem 3.3.1, with probability $1-6\alpha$. The second inequality follows because $\alpha < 1/2$ (note we need $\alpha < 1/18$ to give non-trivial guarantee for the payoff of the mechanism). The third inequality follows from Equation 3.9.16 with probability $1-\alpha$. Looking at the bounds given by the third and the last lines of the above equation (specified by $\star$ and $\star\star$), we get that with probability $1-7\alpha$,

$$\begin{aligned}
I(\mathcal{M}_3) \leq 4\min\left\{\frac{2\ln(1/\alpha)}{\varepsilon} + \sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)}, \frac{4\ln(n/\alpha)}{\varepsilon}\right\} \\
+ \frac{4\sqrt{6}\ln^{1.5}(1/\alpha)}{\varepsilon} + \frac{10\ln(1/\alpha)}{\varepsilon} + \frac{4\ln(2/\alpha)}{3}.
\end{aligned}$$

A similar analysis for $\widetilde{f} \geq 0$ which uses Theorem 3.3.2 gives us the same bound and proves the second part of the theorem. □

3.10 Proof of Theorem 3.3.4

Consider the following family of data sets: first, we initialize D_0 as the data set that has n sellers with valuations $\{1, \ldots, n\}$, and n buyers with valuations $\{n, \ldots, 2n-1\}$.

We then recursively construct D_l for all l. To construct D_{l+1} from D_l, we increase all valuations in D_l by 1, and assign buyers' (resp. sellers) identities in D_{l+1} such that all buyers (resp. sellers) except one have the same valuation as in D_l. Equivalently, our construction works as follows: for any $l \in \mathbb{N}$,

$$D_l = \begin{cases} \mathbf{v}^s = \{l+1, \ldots, l+n\} & \text{sellers' valuations} \\ \mathbf{v}^b = \{l+n, \ldots, l+2n-1\} & \text{buyers' valuations,} \end{cases}$$

up to re-ordering of the agents' identities. The result will follow from the fact that a differentially private algorithm should output similar distributions of prices on data sets D_0 and D_l, but that at the same time, for l large enough, D_0 and D_l are far enough from each other that no distribution of prices can perform well over both of them.

We first show the following lemma, which will be of use in the proof of Theorem 3.3.4.

Lemma 3.10.1. *Let $\{D_l\}$ be the family of data sets described above. If $\mathcal{A} : \mathcal{D}^n \to P$ is an (ε, δ)-DP algorithm, then for every price $p \in P$ and every $k, m \in \mathbb{N}$:*

$$Pr[|\mathcal{A}(D_k) - p| < m] \geq e^{-2k\varepsilon} Pr[|\mathcal{A}(D_0) - p| < m] - 2k\delta.$$

Proof. By the definition of (ε, δ)-DP, if D and D' are neighboring data sets, we must have that for any event E,

$$\Pr[\mathcal{A}(D) \in E] \leq e^{\varepsilon} \Pr[\mathcal{A}(D') \in E] + \delta,$$

or equivalently

$$\Pr[\mathcal{A}(D') \in E] \geq e^{-\varepsilon} \left(\Pr[\mathcal{A}(D) \in E] - \delta\right) \tag{3.10.1}$$

Notice that for every k, by construction, D_k and D_{k+1} differ by only two entries (one buyer's and one seller's valuation). This immediately implies that D_k and D_0

differ by at most $2k$ entries, hence we can apply inequality (3.10.1) recursively $2k$ times to obtain that for any event E,

$$\begin{aligned}\Pr[\mathcal{A}(D_k) \in E] &\geq e^{-\varepsilon}\left(e^{-\varepsilon}\ldots\left(e^{-\varepsilon}\Pr[\mathcal{A}(D_0)\in E]-\delta\right)\ldots-\delta\right)-\delta \\ &= e^{-2k\varepsilon}\Pr[\mathcal{A}(D_0)\in E]-\delta(e^{-(2k-1)\varepsilon}+e^{-(2k-2)\varepsilon}+....+e^{-\varepsilon}+1) \\ &\geq e^{-2k\varepsilon}\Pr[\mathcal{A}(D_0)\in E]-2k\delta\end{aligned}$$

where the last inequality follows from the fact that $e^x \leq 1$ for $x \leq 0$. Fixing the price p and k, m, and taking E to be the ball of radius m around p, i.e.

$$E = \{p' : |p' - p| < m\}$$

concludes the proof. □

We are now ready to prove Theorem 3.3.4.

Proof of Theorem 3.3.4. In this proof, for any given data set $D = (\mathbf{v}^s, \mathbf{v}^b)$, we let

$$u(\mathcal{A}, D) \triangleq \min\left\{\sum_{i\in\mathcal{S}} \mathbb{1}\left[\mathbf{v}_i^s \leq p\right], \sum_{j\in S} \mathbb{1}\left[\mathbf{v}_j^b \geq p\right]\right\}$$

where p is drawn according to $\mathcal{A}(D)$.

First, we note that in data set D_0, at most n trades (where every trading agent gets non-negative utility) can occur, setting a price of n. Further, n is the unique price that makes n trades possible, noting that decreasing (resp. increasing) the price leads to strictly less than n sellers (resp. buyers) willing to trade at that price. We let $p_0^* = n$ be this (unique) optimal price that clears n shares on data set D_0. For a given (ε, δ)-DP algorithm $\mathcal{A} : \mathcal{D}^n \to P$ that outputs a price p given an input data set $D = (\mathbf{v}^s, \mathbf{v}^b)$, let us define, for any $k, m \in \mathbb{N}$ (we will choose these values later on),

$$q_m^0 := \Pr\left[|\mathcal{A}(D_0) - p_0^*| < m\right], \; q_m^k := \Pr\left[|\mathcal{A}(D_k) - p_0^*| < m\right].$$

Notice by Lemma 3.10.1 that

$$q_m^k \geq e^{-2k\varepsilon} q_m^0 - 2k\delta. \tag{3.10.2}$$

Now, fix $m = \lceil \frac{1}{\varepsilon} \rceil$, $k = 2\lceil \frac{1}{\varepsilon} \rceil$, and take $n \geq m$. We have that the expected loss of $\mathcal{A}$ on D_0 is

$$\begin{aligned} \mathrm{E}_{\mathcal{A}}\left[L(\mathcal{A}, D_0)\right] &= \mathrm{opt}(D_0) - \mathrm{E}_{\mathcal{A}}\left[u(\mathcal{A}, D_0)\right] \\ &= n - \mathrm{E}_{\mathcal{A}}\left[u(\mathcal{A}, D_0)\right] \\ &\geq n - \left(q_m^0 \cdot n + (1 - q_m^0) \cdot (n - m)\right) \\ &= (1 - q_m^0) \cdot m \\ &\geq (1 - q_m^0) \cdot \left(\frac{1}{\varepsilon}\right). \end{aligned} \tag{3.10.3}$$

The first inequality follows from a simple application of the law of total expectation on event $E = \{p : |p - p_0^*| < \frac{m}{2n}\}$ and its complement: with probability $1 - q_m^0$ the outputted price is outside E, which implies that it can only clear at most $n - m \geq 0$ shares (picking a price that is m away from n necessarily leads to either m fewer buyers or m fewer sellers willing to trade); the rest of the time, with probability q_m^0, algorithm $\mathcal{A}$ clears at most n shares. The second inequality is an immediate consequence of the choice of m. Similarly, on data set D_k,

$$\begin{aligned} \mathrm{E}_{\mathcal{A}}\left[L(\mathcal{A}, D_k)\right] &= \mathrm{opt}_k - u(\mathcal{A}, D_k) \\ &= n - \mathrm{E}_{\mathcal{A}}\left[u(\mathcal{A}, D_k)\right] \\ &\geq n - \left(q_m^k \cdot (n - (k - m)) + (1 - q_m^k) \cdot n\right) \\ &= q_m^k \cdot (k - m) \\ &\geq \left(e^{-2k\varepsilon} q_m^0 - 2k\delta\right) \cdot (k - m) \\ &\geq \left(e^{-8} q_m^0 - 8(\delta/\varepsilon)\right) \cdot \left(\frac{1}{\varepsilon}\right) \end{aligned} \tag{3.10.4}$$

where the first inequality follows from another use of the law of total expectation on the event E and its complement (notice we choose our parameters so that $k > m$

and $n \geq k-m$): with probability q_m^k, the price is at most $n+m$, and there are $k-m$ sellers that are willing to trade at price $n+m$ but not at price $n+k$, implying that such a price clears at most $n-(k-m)$ shares; the rest of the time, the number of shares cleared is at most n always. The second follows from Equation (3.10.2) and the last one follows from the choice of k and m and the fact that $\varepsilon\lceil\frac{1}{\varepsilon}\rceil \leq 1+\varepsilon \leq 2$ for $0 \leq \varepsilon \leq 1$. Now let $L(\mathcal{A})$ be the worst-case expected loss of $\mathcal{A}$. We have that

$$\begin{aligned} L(\mathcal{A}) &\geq \max\left\{(1-q_m^0), \left(e^{-8}q_m^0 - 8(\delta/\varepsilon)\right)\right\} \cdot \left(\frac{1}{\varepsilon}\right) \\ &\geq \left(\frac{e^{-8} - 8(\delta/\varepsilon)}{1+e^{-8}}\right) \cdot \left(\frac{1}{\varepsilon}\right) \end{aligned}$$

where the first inequality follows from Equations (3.10.3) and (3.10.4) and the second is a simple observation that $f(q_m^0) := \max\{(1-q_m^0), (e^{-8}q_m^0 - 8(\delta/\varepsilon))\}$ is minimized at $q_m^0 = \frac{1+8(\delta/\varepsilon)}{1+e^{-8}}$. Notice the lower bound is valid only when $\delta < \frac{e^{-8}}{8}\varepsilon = \mathcal{O}(\varepsilon)$. This proves our claim that $L(\mathcal{A}) = \Omega\left(\frac{1}{\varepsilon}\right)$. □

3.11 Proofs of Approximate Truthfulness

Our proof of truthfulness for Mechanism 4 will leverage the following lemma, which shows the output of an $(\varepsilon, 0)$-DP mechanism does not change by much in expectation when the input data set is changed by at most one element.

Lemma 3.11.1. *Let $Y = \mathcal{M}(D)$ where $\mathcal{M} : D \to \mathcal{Y}$ is an $(\varepsilon, 0)$-DP mechanism, and let $\max_{y\in\mathcal{Y}} |y| \leq K$. Then for any neighboring data sets $D \sim D'$,*

$$|\,\mathrm{E}\,[Y(D)] - \mathrm{E}\,[Y(D')]\,| \leq (e^{\varepsilon} - 1)K$$

Proof. $Y(D)$ and $Y(D')$ are random variables; we represent the possible values they may take on as $y \in \mathcal{Y}$, and represent the probability distribution of Y under D, D' as $\mathcal{P}$, $\mathcal{P}'$, respectively. It follows that

$$\mathrm{E}\,[Y(D)] - \mathrm{E}\,[Y(D')] = \mathrm{E}_{Y\sim\mathcal{P}}\, Y - \mathrm{E}_{Y\sim\mathcal{P}'}\, Y = \sum_{y\in\mathcal{Y}} (\Pr_{\mathcal{P}}[Y=y] - \Pr_{\mathcal{P}'}[Y=y])\, y$$

Therefore,

$$\begin{aligned}
|\mathrm{E}\left[Y(D)\right] - \mathrm{E}\left[Y(D')\right]| &\leq \sum_{y\in\mathcal{Y}} |\mathrm{Pr}_{\mathcal{P}}[Y=y] - \mathrm{Pr}_{\mathcal{P}'}[Y=y]|\,|y| \\
&\leq \sum_{y\in\mathcal{Y}} (e^{\varepsilon}-1)\max\left\{\mathrm{Pr}_{\mathcal{P}}[Y=y],\ \mathrm{Pr}_{\mathcal{P}'}[Y=y]\right\}|y| \\
&\leq (e^{\varepsilon}-1)K,
\end{aligned}$$

where the second inequality follows from the definition of $(\varepsilon, 0)$-differential privacy.

□

Proof of Claim 9. We prove the claim for any seller. A similar proof holds for buyers. Fix an index i, and any reports/bid vector $\left(\mathbf{r}^s_{-i}, \mathbf{r}^b\right)$ for the remaining buyers and sellers. For simplicity of notation, let us denote $(\mathbf{v}^s_i, \mathbf{r}^s_{-i}, \mathbf{r}^b)$ where i submits his bid truthfully as data set D, and $(\mathbf{r}^s_i, \mathbf{r}^s_{-i}, \mathbf{r}^b)$ for some (other) report $\mathbf{r}^s_i$ as data set D'. Notice D and D' are neighboring data sets. Writing $\mathrm{E}_{\mathcal{M}}$ for the expectation with respect to the mechanism $\mathcal{M}$, we have that:

$$\begin{aligned}
\mathrm{E}_{\mathcal{M}}\left[\mathbf{u}^s_i(\mathcal{M}(D'))\right] &= \mathrm{E}_{\mathcal{M}}[\mathbf{a}^s_i \cdot (p - \mathbf{v}^s_i)|D'] \\
&= \mathrm{E}_{\mathcal{M}}\left[\mathbf{1}\left[p \geq \mathbf{r}^s_i\right] Bern(q^s)(p - \mathbf{v}^s_i)|D'\right] \\
&= \mathrm{E}_{\mathcal{M}}\left[\mathbf{1}\left[p \geq \mathbf{v}^s_i\right] \cdot \mathbf{1}\left[p \geq \mathbf{r}^s_i\right] Bern(q^s)(p - \mathbf{v}^s_i)|D'\right] \\
&+ \mathrm{E}_{\mathcal{M}}\left[\mathbf{1}\left[p < \mathbf{v}^s_i\right] \cdot \mathbf{1}\left[p \geq \mathbf{r}^s_i\right] Bern(q^s)(p - \mathbf{v}^s_i)|D'\right] \\
&\leq \mathrm{E}_{\mathcal{M}}\left[\mathbf{1}\left[p \geq \mathbf{v}^s_i\right] Bern(q^s)(p - \mathbf{v}^s_i)|D'\right] \\
&\leq \mathrm{E}_{\mathcal{M}}\left[\mathbf{1}\left[p \geq \mathbf{v}^s_i\right] Bern(q^s)(p - \mathbf{v}^s_i)|D\right] + \left(e^{3\varepsilon} - 1\right) V \\
&= \mathrm{E}_{\mathcal{M}}\left[\mathbf{u}^s_i(\mathcal{M}(D))\right] + \left(e^{3\varepsilon} - 1\right) V
\end{aligned}$$

where the first inequality follows because the second term appearing in the sum is nonpositive and that $\mathbf{1}\left[p \geq \mathbf{v}^s_i\right] \cdot \mathbf{1}\left[p \geq \mathbf{r}^s_i\right] \leq \mathbf{1}\left[p \geq \mathbf{v}^s_i\right]$. The second inequality follows from Lemma 3.11.1 and the fact that the computation of the pair of random variables (p, q^s) combined with any post-processing of the pair (p, q^s) that is independent of the reported data $D' = (\mathbf{r}^s, \mathbf{r}^b)$ satisfies $(3\varepsilon, 0)$-differential privacy by the

Post-processing Lemma 3.7.3 and the Composition Theorem 3.7.3. Also note that the price/bids range is $\{1, 2, \ldots, V\}$, so we can take $K = V$ in Lemma 3.11.1. □

The proofs of approximate truthfulness of Mechanisms 5 and 6 follow the exact same argument that leverages the stability properties of differential privacy. The only difference comes in the choice of tie-breaking rule and the level of differential privacy of Mechanisms 5 and 6. Rewriting the above proofs with the corresponding tie-breaking rules yields the argument.

3.12 Proofs for Learning Dynamics

3.12.1 Proof of No-Regret Lemma 3.4.4

We first show the claim below:

Claim 10. *Let $R_{j,t}$ be the random variable representing the reward of buyer j in Algorithm 8 at round t, and let $R_j^*(T)$ be the total reward of buyer j's best fixed action in hindsight, over T rounds. Moreover, let $\xi \leq V$ and $\eta \leq \frac{1}{V}$. Then, the regret of buyer j over T rounds is bounded as follows:*

$$R_j^*(T) - \mathrm{E}\left[\sum_{t=1}^{T} R_{j,t}\right] \leq \xi T + \eta V^2 T + \frac{\ln V}{\eta} \tag{3.12.1}$$

Proof. We can think of Algorithm 8 as Exponential Weights with a modified utility function:

$$\text{buyer } j\text{'s modified utility at time } t \text{ for bid } k: \quad \mu_{j,t}^b(k) = \begin{cases} \xi \cdot q_t^b & k = \mathbf{v}_j^b \text{ and } p_t = \mathbf{v}_j^b \\ u_{j,t}(k) & \text{otherwise} \end{cases}$$

where $u_{j,t}(k)$ is the actual utility of buyer j at time t if he were to bid k. Importantly, we show that using this modified utility function we can still achieve vanishing regret (with respect to the *original* reward $R_{j,t}$ which is the agent's *true/realized* utility).

First, notice that $u_{j,t}$ is always upper-bounded by $\mu^b_{j,t}$: $u_{j,t} \leq \mu^b_{j,t}$; but also that $\mu^b_{j,t} \leq u_{j,t} + \xi$. Recall $R^*_j(T)$ is the reward of the best fixed action in hindsight, with respect to the sequence of prices $p_1, \ldots, p_T$ and probabilities $q^b_1, \ldots, q^b_T$ as chosen by an adversary. Let r^*_j be the report that leads to achieving R^*_j, i.e.

$$R^*_j(T) \triangleq \max_{k \in \{1,\ldots,V\}} \sum_{t=1}^{T} u_{j,t}(k), \quad r^*_j \triangleq \text{argmax}_{k \in \{1,\ldots,V\}} \sum_{t=1}^{T} u_{j,t}(k)$$

Our goal will be to show that Equation (3.12.1) holds. Our proof technique will mostly follow standard arguments. In this proof – and this proof only – we let w denote the unnormalized weights that may not sum to 1, and note they induce probability distributions ρ by normalizing each weight by the sum of the weights. First, let $W_{j,t} = \sum_{k=1}^{V} w_{j,t}(k)$. By definition:

$$\frac{W_{j,t+1}}{W_{j,t}} = \frac{\sum_{k=1}^{V} w_{j,t+1}(k)}{\sum_{k=1}^{V} w_{j,t}(k)} = \sum_{k=1}^{V} \frac{w_{j,t}(k) e^{\eta \mu^b_{j,t}(k)}}{\sum_{k=1}^{V} w_{j,t}(k)}$$

We will write $\rho_{j,t}(k) \triangleq w_{j,t}(k) / \sum_{k=1}^{V} w_{j,t}(k)$ as the probability distribution induced by weights $w_{j,t}(k)$, for all k. We can rewrite the above as

$$\frac{W_{j,t+1}}{W_{j,t}} = \sum_{k=1}^{V} \rho_{j,t}(k) e^{\eta \mu^b_{j,t}(k)}.$$

For $\eta \leq \frac{1}{V}$ and $\xi \leq 1$, we have $\eta \mu_{j,t}(k) \leq 1$ for all k. Using the upper bound that $e^x \leq 1 + x + x^2$ for all $x \in [0, 1]$, we obtain that

$$\frac{W_{j,t+1}}{W_{j,t}} \leq 1 + \sum_{k=1}^{V} \rho_{j,t}(k) \cdot \eta \mu^b_{j,t}(k) + \sum_{k=1}^{V} \rho_{j,t}(k) \cdot \eta^2 \mu^b_{j,t}(k)^2$$

Then

$$\begin{aligned} \ln \frac{W_{j,t+1}}{W_{j,t}} &\leq \ln \left(1 + \eta \sum_{k=1}^{V} \rho_{j,t}(k) \mu^b_{j,t}(k) + \eta^2 \sum_{k=1}^{V} \rho_{j,t}(k) \mu^b_{j,t}(k)^2 \right) \\ &\leq \eta \sum_{k=1}^{V} \rho_{j,t}(k) \mu^b_{j,t}(k) + \eta^2 \sum_{k=1}^{V} \rho_{j,t}(k) \mu^b_{j,t}(k)^2, \end{aligned} \tag{3.12.2}$$

where we have used the fact that $\ln(1+x) \leq x$ for $x > -1$ (which holds in this case because payoffs are nonnegative given buyers (sellers) never bid above (below) their valuations). Now noting that $\frac{W_{j,t+1}}{W_{j,1}} = \frac{W_{j,t+1}}{W_{j,t}} \frac{W_{j,t}}{W_{t-1}} \dots \frac{W_{j,2}}{W_{j,1}}$, we can express

$$\ln \frac{W_{j,t+1}}{W_{j,1}} = \ln \frac{W_{j,t+1}}{W_{j,t}} \dots \frac{W_{j,2}}{W_{j,1}} = \sum_{\tau=1}^{t} \ln \frac{W_{\tau+1,j}}{W_{\tau,j}}$$

And applying Inequality (3.12.2), we have that

$$\ln \frac{W_{j,t+1}}{W_{j,1}} \leq \eta \sum_{\tau=1}^{t} \sum_{k=1}^{V} \rho_{j,\tau}(k) \mu^b_{j,\tau}(k) + \eta^2 \sum_{\tau=1}^{t} \sum_{k=1}^{V} \rho_{j,\tau}(k) \mu^b_{j,\tau}(k)^2 \tag{3.12.3}$$

On the other hand, since $W_{j,t+1} \geq w_{j,t}(k)$ for all k, *including* for the best action in hindsight $k = r^*_j$, we have that

$$\begin{aligned}
\ln \frac{W_{j,t+1}}{W_{j,1}} \geq \ln \frac{w_{j,t+1}(r^*_j)}{W_{j,1}} &= \ln(e^{\eta \mu^b_{j,t}(r^*_j)} w_{j,t}(r^*_j)/W_{j,1}) \\
&= \ln(e^{\eta \mu^b_{j,t}(r^*_j)} e^{\eta \mu^b_{j,t-1}(r^*_j)} w_{j,t-1}(r^*_j)) - \ln W_{j,1} \\
&= .... \\
&= \ln \left(\prod_{\tau=1}^{t} e^{\eta \mu^b_{j,\tau}(r^*_j)} w_{j,1}(r^*_j) \right) - \ln W_{j,1}.
\end{aligned}$$

Now, using the fact that the weights can be initialized with $w_{j,1}(k) = 1 \ \forall k$ and $W_{j,1} = V$, this gives

$$\ln \frac{W_{j,t+1}}{W_{j,1}} \geq \eta \sum_{\tau=1}^{t} \mu^b_{j,\tau}(r^*_j) - \ln V \tag{3.12.4}$$

But now combining Inequalities (3.12.3) and (3.12.4) gives:

$$\eta \sum_{\tau=1}^{t} \mu^b_{j,\tau}(r^*_j) - \ln V \leq \eta \sum_{\tau=1}^{t} \sum_{k=1}^{V} \rho_{j,\tau}(k) \mu^b_{j,\tau}(k) + \eta^2 \sum_{\tau=1}^{t} \sum_{k=1}^{V} \rho_{j,\tau}(k) \mu^b_{j,\tau}(k)^2$$

Now notice that $\sum_{k=1}^{V} \rho_{j,\tau}(k) \mu^b_{j,\tau}(k) = \mathrm{E}_k[\mu^b_{j,\tau}(k)]$. So rearranging and letting $t = T$, we have that

$$\sum_{\tau=1}^{T} \mu^b_{j,\tau}(r^*_j) - \sum_{\tau=1}^{T} \mathrm{E}_k[\mu^b_{j,\tau}(k)] \leq \frac{\ln V}{\eta} + \eta \sum_{\tau=1}^{T} \mathrm{E}_k[\mu^b_{j,\tau}(k)^2] \leq \frac{\ln V}{\eta} + \eta T V^2$$

where the inequality follows from the fact that $\mu^b_{j,t}$ is bounded by $\max(V, \xi) = V$ (remembering that $u_{j,t} \leq V$). But since $\mu^b_{j,\tau}(k) \geq u_{j,\tau}(k)$, we have that

$$\begin{aligned} R^*_j(T) - \sum_{\tau=1}^{T} \mathrm{E}_k[\mu^b_{j,\tau}(k)] &= \sum_{\tau=1}^{T} u_{j,\tau}(r^*_j) - \sum_{\tau=1}^{T} \mathrm{E}_k[\mu^b_{j,\tau}(k)] \\ &\leq \sum_{\tau=1}^{T} \mu^b_{j,\tau}(r^*_j) - \sum_{\tau=1}^{T} \mathrm{E}_k[\mu^b_{j,\tau}(k)] \\ &\leq \frac{\ln V}{\eta} + \eta T V^2 \end{aligned}$$

Further, since $\mu^b_{j,\tau}(k) \leq u^b_{j,t}(k) + \xi$, we also have that

$$\begin{aligned} R^*_j(T) - \sum_{\tau=1}^{T} \mathrm{E}[\mu^b_{j,\tau}(k_{j,\tau})] &= \sum_{\tau=1}^{T} u_{j,\tau}(r^*_j) - \sum_{\tau=1}^{T} \mathrm{E}_k[\mu^b_{j,\tau}(k)] \\ &\geq \sum_{\tau=1}^{T} \mu^b_{j,\tau}(r^*_j) - \sum_{\tau=1}^{T} \mathrm{E}_k[\mu^b_{j,\tau}(k)] - \xi T \\ &\geq \sum_{\tau=1}^{T} u_{j,\tau}(r^*_j) - \sum_{\tau=1}^{T} \mathrm{E}_k[u_{j,\tau}(k)] - \xi T \\ &= R^*_j(T) - \sum_{t=1}^{T} \mathrm{E}[R_{j,t}] - \xi T. \end{aligned}$$

Combining the last two inequalities, we get

$$R^*_j\left(T\right) - \sum_{t=1}^{T} \mathrm{E}\left[R_{j,t}\right] \leq \xi T + \frac{\ln V}{\eta} + \eta V^2 T,$$

as desired. □

We can now conclude the proof, noting that Lemma 10 gives that the total regret of Algorithm 8 over T rounds for agent j is bounded by:

$$\text{Regret} \leq \xi T + \frac{\ln V}{\eta} + \eta V^2 T$$

Choose $\eta = \frac{1}{V\sqrt{T}}$ and $\xi = \frac{1}{\sqrt{T}}$. Then we have that

$$\text{Regret} \leq \sqrt{T} + V \ln V \sqrt{T} + \frac{1}{V\sqrt{T}} V^2 T = \sqrt{T} + V \ln V \sqrt{T} + V\sqrt{T}.$$

Then average regret can be bounded as:

$$\frac{1}{T}\,\text{Regret} \leq \frac{1}{\sqrt{T}} + \frac{V \ln V}{\sqrt{T}} + \frac{V}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

That is, average regret vanishes as $T \to \infty$.

3.12.2 Proof of Theorem 3.4.1

To prove Theorem 3.4.1, we will examine how the opt$'$ sellers with the lowest values and the opt$'$ buyers with the highest values update their weights. To do so, we will need the following definition:

Definition 3.12.1 (Highest (resp. lowest) value buyers (resp. sellers))**.** Let $n^b(v) = \sum_{i=1}^{n^b} \mathbf{1}[\mathbf{v}_j^b \geq v]$ be the number of buyers with value bigger than or equal to v, and let $\nu^b = \max\{v : \ n^b(v) \geq \text{opt}'\}$. Similarly, let $n^s(v) = \sum_{i=1}^{n^s} \mathbf{1}[\mathbf{v}_j^b \leq v]$ be the number of sellers with value smaller than or equal to v, and let $\nu^s = \min\{v : \ n^s(v) \geq \text{opt}'\}$.

We note the following property of ν^b, ν^s:

Claim 11. *Suppose opt$'$ > 0. Then,*

$$\nu^b \geq \nu^s + 2.$$

Proof. By definition of opt$'$, there exists a price $p^\star$ such that at least opt$'$ buyers have value above or equal to $p^\star + 1$ and opt$'$ sellers below or equal to $p^\star - 1$. But then, $\nu^s \leq p^\star - 1$ and $\nu^b \geq p^\star + 1$, which concludes the proof. □

First of all, we show that if a given price p is picked infinitely many times, every agent j with $\mathbf{v}_j^b > p$ sees their probability of bidding more than p converge to 1. This is the object of Corollary 2, whose proof relies on Lemmas 3.12.2 and 3.12.3 below. We state the Lemmas for a buyer j and note that similar results hold for a seller i as well.

Lemma 3.12.2. *For all t, for all $p \in [V]$, for all $j \in [n^b]$,*

$$\sum_{k=p}^{\mathbf{v}_j^b} w_{j,t+1}^b(k) \geq \sum_{k=p}^{\mathbf{v}_j^b} w_{j,t}^b(k).$$

Proof. If $\sum_{k<p} w_{j,t}^b(k) = 0$, the result is immediate: it must be that for all $k < p$, $w_{j,t}^b(k) = 0$, so by exponential update, $w_{j,t+1}^b(k) = 0$, leading to $\sum_{k<p} w_{i,t+1}^b(k) = 0$. In turn,

$$\sum_{k\geq p} w_{j,t+1}^b(k) = \sum_{k\geq p} w_{j,t}^b(k) = 1.$$

We now focus on the case when $\sum_{k<p} w_{j,t}^b(k) > 0$. Remember that p_t is the optimal price at time t. If $p \leq p_t$,

$$\frac{\sum_{k\geq p} w_{j,t+1}^b(k)}{\sum_{k<p} w_{j,t+1}^b(k)} = \frac{\sum_{k=p}^{p_t-1} w_{j,t}^b(k) + \sum_{k\geq p_t} w_{j,t}^b(k)\exp(\eta q_t^b(\mathbf{v}_j^b - p_t))}{\sum_{k<p} w_{j,t}^b(k)} \geq \frac{\sum_{k\geq p} w_{j,t}^b(k)}{\sum_{k<p} w_{j,t}^b(k)}.$$

When $p > p_t$,

$$\begin{aligned}\frac{\sum_{k\geq p} w_{j,t+1}^b(k)}{\sum_{k<p} w_{j,t+1}^b(k)} &= \frac{\sum_{k\geq p} w_{j,t}^b(k)\exp(\eta q_t^b(\mathbf{v}_j^b - p_t))}{\sum_{k<p_t} w_{j,t}^b(k) + \sum_{k=p_t}^{p-1} w_{j,t}^b(k)\exp(\eta q_t^b(\mathbf{v}_j^b - p_t))} \\ &\geq \frac{\sum_{k\geq p} w_{j,t}^b(k)\exp(\eta q_t^b(\mathbf{v}_j^b - p_t))}{\left(\sum_{k<p_t} w_{j,t}^b(k) + \sum_{k=p_t}^{p-1} w_{j,t}^b(k)\right)\exp(\eta q_t^b(\mathbf{v}_j^b - p_t))} \\ &= \frac{\sum_{k\geq p} w_{j,t}^b(k)}{\sum_{k<p} w_{j,t}^b(k)}.\end{aligned}$$

Since

$$\sum_{k\geq p} w_{j,t+1}^b(k) + \sum_{k<p} w_{j,t+1}^b(k) = 1, \sum_{k\geq p} w_{j,t}^b(k) + \sum_{k<p} w_{j,t}^b(k) = 1,$$

we have that for all p,

$$\frac{\sum_{k\geq p} w_{j,t+1}^b(k)}{1 - \sum_{k\geq p} w_{j,t+1}^b(k)} \geq \frac{\sum_{k\geq p} w_{j,t}^b(k)}{1 - \sum_{k\geq p} w_{j,t}^b(k)}.$$

This in particular implies that for all p,

$$\sum_{k\geq p} w_{j,t+1}^b(k)\left(1 - \sum_{k\geq p} w_{j,t}^b(k)\right) \geq \sum_{k\geq p} w_{j,t}^b(k)\left(1 - \sum_{k\geq p} w_{j,t+1}^b(k)\right),$$

hence

$$\sum_{k \geq p} w^b_{j,t+1}(k) \geq \sum_{k \geq p} w^b_{j,t}(k).$$

□

Lemma 3.12.3 (Update moves mass up by a constant amount)**.** *Suppose at time t, at least one buyer and one seller can trade. There exists a constant $C(\varepsilon) > 1$ such that for any buyer j with $\mathbf{v}^b_j > p_t$ and $\sum_{k=p_t}^{\mathbf{v}^b_j} w^b_{j,t}(k) \leq 1 - \varepsilon$, we have that*

$$\frac{\sum_{k=p_t}^{\mathbf{v}^b_j} w^b_{j,t+1}(k)}{\sum_{k=p_t}^{\mathbf{v}^b_j} w^b_{j,t}(k)} \geq C(\varepsilon).$$

Proof. Let $X_t(p)$ be the probability that buyer j bids at least p on round t. For simplicity of notations, we omit the j subscripts in the proof. Trivially:

$$X_t(p_t) = \sum_{k=p_t}^{\mathbf{v}^b_j} w^b_{j,t}(k).$$

Now, by the definition of exponential weights, we have that

$$X^{t+1}(p_t) = \frac{e^{\eta q^b_t(\mathbf{v}^b_j - p_t)} X_t(p_t)}{e^{\eta q^b_t(\mathbf{v}^b_j - p_t)} X_t(p_t) + (1 - X_t(p_t))}$$

since the buyer updates $w^b_{j,t}(k)$ with $e^{\eta q^b_t(\mathbf{v}^b_j - p_t)}$ for all bids k above p_t up to $\mathbf{v}^b_j$, and updates weights on bids $k \leq p_t$ with $e^{\eta q^b_t \cdot 0} = 1$. It immediately follows that

$$\frac{X^{t+1}(p_t)}{X_t(p_t)} = \frac{e^{\eta q^b_t(\mathbf{v}^b_j - p_t)}}{X_t(p_t)(e^{\eta q^b_t(\mathbf{v}^b_j - p_t)} - 1) + 1}$$

Now by assumption, $X_t(p_t) < 1 - \varepsilon$, so

$$\begin{aligned}
\frac{X^{t+1}(p_t)}{X_t(p_t)} &> \frac{e^{\eta q^b_t(\mathbf{v}^b_j - p_t)}}{(1-\varepsilon)(e^{\eta q^b_t(\mathbf{v}^b_j - p_t)} - 1) + 1} \\
&= \frac{e^{\eta q^b_t(\mathbf{v}^b_j - p_t)}}{e^{\eta q^b_t(\mathbf{v}^b_j - p_t)} - \varepsilon e^{\eta q^b_t(\mathbf{v}^b_j - p_t)} + \varepsilon} \\
&= \frac{e^{\eta q^b_t(\mathbf{v}^b_j - p_t)}}{e^{\eta q^b_t(\mathbf{v}^b_j - p_t)} + \varepsilon(1 - e^{\eta q^b_t(\mathbf{v}^b_j - p_t)})} \\
&= \frac{1}{1 - \varepsilon(1 - \frac{1}{e^{\eta q^b_t(\mathbf{v}^b_j - p_t)}})}
\end{aligned}$$

Using the fact that $q_t^b \geq \frac{1}{n^b}$, as there are at most n^b buyers and at least one possible seller to trade with, and the fact that $\mathbf{v}_j^b - p_t \geq 1$, we get that

$$e^{\eta/n^b} \leq e^{\eta q_t^b (\mathbf{v}_j^b - p_t)}.$$

In turn,

$$\frac{X^{t+1}(p_t)}{X_t(p_t)} \geq \frac{1}{1 - \varepsilon(1 - e^{-\eta/n^b})} > 1.$$

Letting $C(\varepsilon) = \frac{1}{1-\varepsilon(1-e^{-\eta/n^b})}$ is enough to conclude the proof. □

Corollary 2. *Pick any buyer j, and let $p < \mathbf{v}_j^b$. Let $N_t(p)$ be the number of times price p is picked by the mechanism so that at least one trade is possible at p, up until time t. In other words,*

$$N_t(p) = \sum_{t' \leq t} \mathbf{1}\left[\Pi_{t'}\left(p, \mathbf{r}_{t'}^s, \mathbf{r}_{t'}^b\right) \geq 1\right]$$

If $\lim_{t\to\infty} N_t(p) = +\infty$, then

$$\lim_{t\to\infty} Pr[\mathbf{r}_{j,t}^b \geq p] = 1$$

Proof. Fix $\varepsilon > 0$. At time t, by applying Lemma 3.12.3 and Lemma 3.12.2 repeatedly, we have that

$$\begin{aligned}\Pr[\mathbf{r}_{j,t}^b \geq p] &\geq \min\{1-\varepsilon, C(\varepsilon)^{N_t(p)} \Pr[\mathbf{r}_{j,0}^b \geq p]\} \\ &\geq \min\{1-\varepsilon, C(\varepsilon)^{N_t(p)} \frac{1}{V}\}\end{aligned}$$

where the last inequality follows because the initial weights are uniform over all bids. By assumption, there exists T such that for all $t \geq T$: $N_t(p) \geq \frac{\log((1-\varepsilon)(V))}{\log C(\varepsilon)}$, and consequently,

$$\Pr[\mathbf{r}_{j,t}^b \geq p] \geq 1 - \varepsilon.$$

Since this holds for every $\varepsilon > 0$, the limit statement follows. □

We note that a similar Corollary exists for sellers as well. Now, we need to show that there is a price that clears benchmark opt′ and is chosen by the mechanism infinitely often. This is the object of Lemma 3.12.4, whose proof relies on Claim 12. Once again, we state the Claim only for buyers and note that a similar result for sellers as well.

Claim 12. *For every buyer j, for all t, $\frac{1}{V} \leq w^b_{j,t}(\mathbf{v}^b_j) \leq \frac{1}{2}$.*

Proof. At time step t, if $p_t > \mathbf{v}^b_j$, agent j does not update any weight. If $p_t \leq \mathbf{v}^b_j$, it is easy to see that the weight on $\mathbf{v}^b_j$ cannot decrease in the next round. Indeed, for any k such that $p_t \leq k \leq \mathbf{v}^b_j$, we have that

$$\begin{aligned}
w^b_{j,t+1}(k) &= w^b_{j,t}(k) \cdot \frac{\exp\left(\eta q^b_t(\mathbf{v}^b_j - p_t)\right)}{\sum_{k<p_t} w^b_{j,t}(k) + \sum_{k\geq p_t} w^b_{j,t}(k)\exp\left(\eta q^b_t(\mathbf{v}^b_j - p_t)\right)} \\
&= w^b_{j,t}(k) \cdot \frac{1}{\exp\left(-\eta q^b_t(\mathbf{v}^b_j - p_t)\right)\sum_{k<p_t} w^b_{j,t}(k) + \sum_{k\geq p_t} w^b_{j,t}(k)} \\
&= w^b_{j,t}(k) \cdot \frac{1}{\exp\left(-\eta q^b_t(\mathbf{v}^b_j - p_t)\right)\sum_{k<p_t} w^b_{j,t}(k) + 1 - \sum_{k<p_t} w^b_{j,t}(k)} \\
&= w^b_{j,t}(k) \cdot \frac{1}{1 - \left(1 - \exp\left(-\eta q^b_t(\mathbf{v}^b_j - p_t)\right)\right)\sum_{k<p_t} w^b_{j,t}(k)} \\
&\geq w^b_{j,t}(k),
\end{aligned}$$

where the last step follows from noting that both $1-\exp\left(-\eta q^b_t(\mathbf{v}^b_j - p_t)\right)$, $\sum_{k<p_t} w^b_{j,t}(k) \leq 1$. As such, $w^b_{j,t}(\mathbf{v}^b_j)$ is non-decreasing in t, so $w^b_{j,t}(\mathbf{v}^b_j) \geq w^b_{j,0}(\mathbf{v}^b_j) = \frac{1}{V}$.

Let us now prove the second inequality. Note that at any time step t, let p_t be the price chosen by the mechanism. When $p_t > \mathbf{v}^b_j$, j does not update his weight. Similarly, when $p_t = \mathbf{v}^b_j$, the exponential update rule is the same for $w^b_{j,t}(\mathbf{v}^b_j)$ and $w^b_{j,t}(\mathbf{v}^b_j - 1)$ and given by $\exp\left(\eta q^b_t(\mathbf{v}^b_j - p_t)\right) = \exp(0) = 1$. When $p_t < \mathbf{v}^b_j$, both $w^b_{j,t}(\mathbf{v}^b_j)$ and $w^b_{j,t}(\mathbf{v}^b_j - 1)$ are multiplied by the same amount $\exp\left(\eta q^b_t(\mathbf{v}^b_j - p_t)\right)$. Therefore, it immediately follows by induction that $w^b_{j,t}(\mathbf{v}^b_j) = w^b_{j,t}(\mathbf{v}^b_j - 1)$ for all t. In particular, this implies $w^b_{j,t}(\mathbf{v}^b_j) \leq 1/2$, as $w^b_{j,t}(\mathbf{v}^b_j) + w^b_{j,t}(\mathbf{v}^b_j - 1) \leq 1$. □

Lemma 3.12.4 (Good event)**.** *Suppose* $\mathrm{opt}' > 0$*, and let*

$$\gamma \triangleq \left(\frac{1}{V}\right)^{1+|n^b(\nu^s+1)||n^s(\nu^b-1)|} \cdot \left(\frac{1}{2}\right)^{\left(n^b-|n^b(\nu^s+1)|\right)\left(n^s-|n^s(\nu^b-1)|\right)} > 0.$$

At any time t*,* $\nu^s < p_t < \nu^b$ *and at least one trade is possible with probability at least* γ*.*

Proof. By Claim 12, we have that with probability at least

$$\left(\frac{1}{V}\right)^{|n^b(\nu^s+1)||n^s(\nu^b-1)|} \cdot \left(\frac{1}{2}\right)^{\left(n^b-|n^b(\nu^s+1)|\right)\left(n^s-|n^s(\nu^b-1)|\right)} = V\gamma,$$

all buyers with value $\mathbf{v}_j^b > \nu^s$ bid their value, all buyers with value $\mathbf{v}_j^b \leq \nu^s$ bid strictly below their value, all sellers with value $\mathbf{v}_i^s < \nu^b$ bid their value, and all sellers with $\mathbf{v}_i^s \geq \nu^b$ bid strictly more than their value. In particular, since $\nu^s < \nu^b$, all buyers with value $\mathbf{v}_j^b \geq \nu^b$ bid their value and all sellers with value $\mathbf{v}_i^s \leq \nu^s$ bid their value. By definition of ν^b and ν^s, there are at least opt' such buyers and sellers, so setting any price p satisfying $\nu^s \leq p \leq \nu^b$ clears opt' shares at least. On the other hand, any price $p > \nu^b$ and any price $p < \nu^s$ cannot clear opt' shares. Therefore, $\nu^s \leq p_t \leq \nu^b$. Further, since all buyers with value $\mathbf{v}_j^b \geq \nu^b$ and all sellers with value $\mathbf{v}_i^s \leq \nu^s$ bid their values, and $\nu^b \geq \nu^s$, at least $\mathrm{opt}' \geq 1$ trades happen at price p_t.

When $\nu^s < p < \nu^b$ for all optimal prices, this is enough to conclude the proof. Now, suppose $p = \nu^b$ is an optimal price at time t. By construction, no seller bids ν^b. As such, the number of sellers with bids under p and the number of sellers with bids under $p-1$ are the same, and $p - 1 = \nu^b - 1$ clears at least as many shares as p, hence is optimal at time t. Because p_t is chosen uniformely at random among the set of optimal prices, and there are at most V optimal prices, $p-1$ is picked with probability at least $\frac{1}{V}$, and satisfies $\nu^s < p - 1 < \nu^b$ by Claim 11. Similarly, if $p = \nu^s$ is optimal, then so is $p + 1 < \nu^b$, and it is picked by the mechanism with probability at least $\frac{1}{V}$. This concludes the proof. □

We are now ready to put everything together, and show Theorem 3.4.1.

Proof of Theorem 3.4.1. The case when $\text{opt}' = 0$ is immediate. So let us assume $\text{opt}' > 0$. Lemma 3.12.4 shows that at any given round, there is a constant probability $\gamma > 0$ to pick $p_t \in (\nu^s, \nu^b)$ and realize at least one trade at that price. As such, as $t \to +\infty$, the number of times the mechanism picks a price in (ν^s, ν^b) such that a trade is realized also tends to infinity. In particular, by the pigeonhole principle, there exists a price $p^\star \in (\nu^s, \nu^b)$ such that

$$\lim_{t\to\infty} N_t(p^\star) = +\infty.$$

By Corollary 2, for every buyer $j \in n^b(\nu^b)$,

$$\lim_{t\to\infty} \Pr[\mathbf{r}^b_{j,t} \geq p^\star] = 1,$$

and similarly, for every seller $i \in n^s(\nu^s)$,

$$\lim_{t\to\infty} \Pr[\mathbf{r}^s_{i,t} \leq p^\star] = 1.$$

Since there are at least opt' buyers in $n^b(\nu^b)$ and opt' sellers in $n^s(\nu^s)$, we have that

$$1 \geq \Pr\left[\Pi_t\left(p_t, \mathbf{r}^s, \mathbf{r}^b\right) \geq \text{opt}'\right] \geq \prod_{j\in n^b(\nu^b)} \Pr[\mathbf{r}^b_{j,t} \geq p^\star] \cdot \prod_{i\in n^s(\nu^s)} \Pr[\mathbf{r}^s_{i,t} \leq p^\star],$$

which concludes the proof. □

3.12.3 Proof of Theorem 3.4.2

The proof is similar to that of Theorem 3.4.1, and is given below. We start by showing in Corollary 3 that if a price p is picked by the mechanism infinitely many times, every buyer with value at least p learns to bid higher than p with probability going to 1.

Lemma 3.12.5. *For all t, for all $p \in [V]$, for all buyers j,*

$$\sum_{k=p}^{\mathbf{v}^b_j} \mathbf{w}^b_{j,t+1}(k) \geq \sum_{k=p}^{\mathbf{v}^b_j} \mathbf{w}^b_{j,t}(k).$$

Proof. The proof is identical to that of Lemma 3.12.2. □

We then characterize by how much the weight allocated to bids above the chosen price p_t increase for a buyer j, at every time step t:

Lemma 3.12.6 (Update moves mass up by a constant amount)**.** *Suppose at time t, at least one buyer and one seller can trade. There exists a constant $C(\varepsilon) > 1$ such that for any buyer j with $\mathbf{v}_j^b \geq p_t$ and $\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t}^b(k) \leq 1 - \varepsilon$, we have that*

$$\frac{\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t+1}^b(k)}{\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t}^b(k)} \geq C(\varepsilon).$$

Proof. Note that when $p_t < \mathbf{v}_j^b$, we have by Lemma 3.12.3 that

$$\frac{\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t+1}^b(k)}{\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t}^b(k)} \geq \frac{1}{1 - \varepsilon(1 - e^{-\eta/n^b})} > 1.$$

Now, when $p_t = \mathbf{v}_j^b$ note that

$$\frac{\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t+1}^b(k)}{\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t}^b(k)} = \frac{\mathbf{w}_{j,t+1}^b(\mathbf{v}_j^b)}{\mathbf{w}_{j,t}^b(\mathbf{v}_j^b)} = \exp\left(\eta q_t^b \xi\right).$$

In particular, as there is at least one possible trade, we have that $q_t^b \geq 1/n^b$, hence

$$\frac{\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t+1}^b(k)}{\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t}^b(k)} \geq \exp\left(\frac{\eta\xi}{n^b}\right).$$

Letting $C(\varepsilon) = \min\left(\frac{1}{1-\varepsilon(1-e^{-\eta/n^b})}, \exp\left(\frac{\eta\xi}{n^b}\right)\right)$ is enough to conclude the proof. □

Corollary 3. *Pick any buyer j, and let $p \leq \mathbf{v}_j^b$. Let $N_t(p)$ be the number of times price p is picked and at least one trade is possible at price p, up until time t. If $\lim_{t\to\infty} N_t(p) = +\infty$, then*

$$\lim_{t\to\infty} Pr[\mathbf{r}_{j,t}^b \geq p] = 1$$

Proof. This is identical to the proof of Corollary 2. □

Second, we need to show that there is a price that clears benchmark opt$'$ and is chosen by the mechanism infinitely often.

Lemma 3.12.7 (Good event)**.** *With probability at least* $\left(\frac{1}{V}\right)^{n^b+n^s}$*, all agents bid their valuation.*

Proof. By the same proof as Corollary 12, for every agent j and for all t, $\frac{1}{V} \leq \mathbf{w}^b_{j,t}(\mathbf{v}^b_j)$. This is enough to prove the lemma. □

We are now ready to put everything together, and show Theorem 3.4.2.

Proof of Theorem 3.4.2. Suppose opt > 0 (otherwise the result is immediate). When all agents bid their values, the mechanism selects a price that executes opt ≥ 1 trades. Lemma 3.12.7 shows this happens with constant probability at any given round, and as such happens infinitely often when the number of rounds goes to infinity. By the pigeonhole principle, there exists a price $p^\star$ such that there are at least opt buyers (resp. sellers) with value at least (resp. at most) $p^\star$, and such that

$$\lim_{t\to\infty} N_t(p^\star) = +\infty.$$

By Corollary 3, for any buyer with $\mathbf{v}^b_j \geq p^\star$,

$$\lim_{t\to\infty} \Pr[\mathbf{r}^b_{j,t} \geq p^\star] = 1,$$

and similarly, for every seller i with $\mathbf{v}^s_i \leq p^\star$,

$$\lim_{t\to\infty} \Pr[\mathbf{r}^s_{i,t} \leq p^\star] = 1.$$

In turn, since

$$1 \geq \Pr\left[\Pi_t\left(p_t, \mathbf{r}^s_t, \mathbf{r}^b_t\right) \geq \text{opt}\right] \geq \prod_{j\in[n^b]:\ \mathbf{v}^b_j \geq p^\star} \Pr[\mathbf{r}^b_{j,t} \geq p^\star] \cdot \prod_{i\in[n^s]:\ \mathbf{v}^s_i \leq p^\star} \Pr[\mathbf{r}^s_{i,t} \leq p^\star],$$

we have

$$\lim_{t\to+\infty} \Pr\left[\Pi_t\left(p_t, \mathbf{r}^s_t, \mathbf{r}^b_t\right) \geq \text{opt}\right] = 1.$$

□

3.12.4 Proof of Theorem 3.4.3

We start by noting that in the private case, the weights are still non-decreasing over time.

Lemma 3.12.8. *For all t, for all $p \in [V]$, for all buyers j,*

$$\sum_{k=p}^{\mathbf{v}_j^b} \mathbf{w}_{j,t+1}^b(k) \geq \sum_{k=p}^{\mathbf{v}_j^b} \mathbf{w}_{j,t}^b(k).$$

One distinction compared to the non-private case arises with respect to the amount by which the weights above p_t are updated. This amount depends on q_t^b, which is a random variable over the randomness of private computation of the selection probability. We note that *conditionally on* $q_t^b \geq 1/n^b$, Lemma 3.12.3 carries through, as formalized below:

Lemma 3.12.9 (Update moves mass up by a constant amount)**.** *Suppose at time t, at least one buyer and one seller can trade and that $q_t^b > 1/n^b$. There exists a constant $C(\varepsilon) > 1$ such that for any buyer j with $\mathbf{v}_j^b \geq p_t$ and $\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t}^b(k) \leq 1-\varepsilon$,*

$$\frac{\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t+1}^b(k)}{\sum_{k=p_t}^{\mathbf{v}_j^b} \mathbf{w}_{j,t}^b(k)} \geq C(\varepsilon).$$

We now fix a price p. We show that for one such p, if the event where p is the price picked by the mechanism *and* $q_t^b \geq 1/n$ happens infinitely often, then all bidders with valuation equal to or larger than p learn to bid higher than p with probability that goes to 1.

Lemma 3.12.10. *Pick any buyer j, and let $p \leq \mathbf{v}_j^b$. Let $N_t(p)$ be the number of times price p is picked by the mechanism so that at least one trade is possible at p and $q^b > 1/n^b$, up until time t. In other words,*

$$N_t(p) = \sum_{t' \leq t} \mathbf{1}\left[\Pi_{t'}\left(p, \mathbf{r}_{t'}^s, \mathbf{r}_{t'}^b\right) \geq 1, q_{t'}^b > \frac{1}{n^b}\right]$$

If $\lim_{t\to\infty} N_t(p) = +\infty$, then $\lim_{t\to\infty} Pr[\mathbf{r}_{j,t}^b \geq p] = 1$.

We note that the event in which all agents bid their valuation and the mechanism (despite the randomness due to privacy) picks an optimal price p and releases $q^b \geq 1/n^b$, $q^s \geq 1/n^s$ happens with at least constant probability (independent of the time dimension of the problem), hence infinitely many times when the time horizon goes to infinity:

Lemma 3.12.11 (Good event)**.** *Suppose opt ≥ 1. At any round t, with probability at least $C\left(\frac{1}{V}\right)^{n^b+n^s+1}$ for some constant $C > 0$: all buyers bid their valuations, $q_t^b > 1/n^b$, $q_t^s > 1/n^s$, and the chosen price p_t is an optimal price that clears OPT shares.*

Proof. We have shown before in the proof of Lemma 3.12.7 that with probability at least $V^{-(n^b+n^s)}$ every agent bids their valuation.

In the rest of the proof, we condition on all agents bidding their valuations in the current round t. Conditional on this, we show that with constant probability, simultaneously: $q_t^b > 1/n^b$, and $q_t^s > 1/n^s$. Recall from Algorithm 4 that in each round t, given the selected price p_t, we have

$$q_t^b = \min\left(1, \frac{(\widehat{s}_t)_+}{\left(\widehat{b}_t - \frac{\ln(1/\alpha)}{\varepsilon}\right)_+}\right), \quad q_t^s = \min\left(1, \frac{\left(\widehat{b}_t\right)_+}{\left(\widehat{s}_t - \frac{\ln(1/\alpha)}{\varepsilon}\right)_+}\right).$$

By the accuracy guarantees of the Laplace mechanism and the fact that Laplace noise has positive value with probability $1/2$, we have that with constant probability C (for some C that only depends on α and ε but not on t), the 4 following events simultaneously hold:

1. $\left(\widehat{b}_t - \frac{\ln(1/\alpha)}{\varepsilon}\right)_+ \leq \sum_{j\in\mathcal{B}} \mathbb{1}\left[\mathbf{v}_j^b \geq p_t\right] \leq n^b$,
2. $\widehat{b}_t \geq \sum_{j\in B} \mathbb{1}\left[\mathbf{v}_j^b \leq p_t\right] \geq \text{opt} \geq 1$,
3. $\left(\widehat{s}_t - \frac{\ln(1/\alpha)}{\varepsilon}\right)_+ \leq \sum_{i\in\mathcal{S}} \mathbb{1}\left[\mathbf{v}_i^s \leq p_t\right] \leq n^s$,
4. $\widehat{s}_t \geq \sum_{i\in\mathcal{S}} \mathbb{1}\left[\mathbf{v}_i^s \leq p_t\right] \geq \text{opt} \geq 1$, noting that at least one trade is possible at price p_t.

Using the above inequalities, we obtain that with probability C,

$$q_t^b = \min\left(1, \frac{(\widehat{s}_t)_+}{\left(\widehat{b}_t - \frac{\ln(1/\alpha)}{\varepsilon}\right)_+}\right) \geq \min\left(1, \frac{1}{n^b}\right) \geq \frac{1}{n^b},$$

$$q_t^s = \min\left(1, \frac{\left(\widehat{b}_t\right)_+}{\left(\widehat{s}_t - \frac{\ln(1/\alpha)}{\varepsilon}\right)_+}\right) \geq \min\left(1, \frac{1}{n^s}\right) \geq \frac{1}{n^s}.$$

To finish the proof, we just need to show that, conditional on all agents bidding their valuations in the current round t, with probability at least $1/V$, p_t – the price selected when every agent bids their valuation – is an optimal price. Note there exists a price $p_t^\star$ that is optimal for round t, i.e. such that $\Pi_t(p_t^\star, \mathbf{v}^s, \mathbf{v}^b) \geq \Pi_t(p, \mathbf{v}^s, \mathbf{v}^b)$ for all p. By the exponential mechanism, this price $p_t^\star$ is selected with probability

$$\frac{\exp\left(\varepsilon\Pi_t(p_t^\star, \mathbf{v}^s, \mathbf{v}^b)/2\right)}{\sum_{p=1}^V \exp\left(\varepsilon\Pi_t(p, \mathbf{v}^s, \mathbf{v}^b)/2\right)} \geq \frac{\exp\left(\varepsilon\Pi_t(p_t^\star, \mathbf{v}^s, \mathbf{v}^b)/2\right)}{\sum_{p=1}^V \exp\left(\varepsilon\Pi_t(p_t^\star, \mathbf{v}^s, \mathbf{v}^b)/2\right)} = \frac{1}{V}.$$

□

We are now ready to put everything together, and show Theorem 3.4.3.

Proof of Theorem 3.4.3. Let us for simplicity call:

$$f(\varepsilon, \alpha) \triangleq \frac{2\ln(V/\alpha)}{\varepsilon} - \frac{2\ln(1/\alpha)}{\varepsilon} - \sqrt{6\left(\text{opt} + \frac{\ln(1/\alpha)}{\varepsilon}\right)\ln(1/\alpha)}$$

Suppose opt > 0 (otherwise the result is immediate). By Lemma 3.12.11 that shows that with constant probability (independent of time) in every round, the mechanism picks an optimal price and $q^b, q^s \geq \frac{1}{n}$, this event must happen infinitely many times. By the pigeonhole principle, there exists an optimal price $p^\star$ such that infinitely many times, $p^\star$ is picked by the mechanism with $q^b, q^s \geq \frac{1}{n}$. In turn, all buyers j with $\mathbf{v}_j^b \geq p^\star$ and all sellers i with $\mathbf{v}_i^s \leq p^\star$ (there are at least OPT of them, since $p^\star$ is optimal) learn to bid above, respectively below price $p^\star$ with probability

that tends to 1 as t goes to infinity, by Lemma 3.12.10. Formally, for every buyer j with $\mathbf{v}_j^b \geq p^\star$,

$$\lim_{t\to\infty} \Pr[\mathbf{r}_{j,t}^b \geq p^\star] = 1,$$

and similarly, for every seller i with $\mathbf{v}_i^s \leq p^\star$, we have that

$$\lim_{t\to\infty} \Pr[\mathbf{r}_{i,t}^s \leq p^\star] = 1.$$

Hence

$$\lim_{t\to\infty} \prod_{j\in[n^b]:\ \mathbf{v}_j^b \geq p^\star} \Pr[\mathbf{r}_{j,t}^b \geq p^\star] \cdot \prod_{i\in[n^s]:\ \mathbf{v}_i^s \leq p^\star} \Pr[\mathbf{r}_{i,t}^s \leq p^\star] = 1$$

and consequently, there exists $N(\alpha)$ large enough such that for all $t \geq N(\alpha)$,

$$\prod_{j\in[n^b]:\ \mathbf{v}_j^b \geq p^\star} \Pr[\mathbf{r}_{j,t}^b \geq p^\star] \cdot \prod_{i\in[n^s]:\ \mathbf{v}_i^s \leq p^\star} \Pr[\mathbf{r}_{i,t}^s \leq p^\star] \geq 1 - \alpha.$$

When all buyers with value at least the price and sellers with value at most the price bid between their valuation and $p^\star$, the optimal number of shares that can be cleared is OPT. By the accuracy guarantee of Mechanism 4, it must then be the case that for all $t \geq N(\alpha)$,

$$\begin{aligned}
\Pr\left[\Pi_t(p_t, \mathbf{r}_t^s, \mathbf{r}_t^b) \geq \text{opt} - f(\varepsilon, \alpha)\right] &\geq (1 - 8\alpha) \prod_{j\in[n^b]:\ \mathbf{v}_j^b \geq p^\star} \Pr[\mathbf{r}_{j,t}^b \geq p^\star] \cdot \prod_{i\in[n^s]:\ \mathbf{v}_i^s \leq p^\star} \Pr[\mathbf{r}_{i,t}^s \leq p^\star] \\
&\geq (1 - 8\alpha)(1 - \alpha) \\
&\geq 1 - 9\alpha.
\end{aligned}$$

This concludes the proof. □

The proof for benchmark opt′ follows the same argument, and is omitted for simplicity of exposition.

Part II

MARKETS WITH ALGORITHMS

Chapter 4

COMPETITION, REGULATION, AND ERROR INEQUALITY IN DATA-DRIVEN MARKETS

4.1 Introduction[1]

As machine learning has become more integrated into products, markets, and decision-making throughout society, researchers, practitioners, and activists have identified many instances of *unfairness* in predictions or decisions made by machine-learned models (or by humans influenced by said models). A large and developing body of work, which we briefly survey in Section 4.2, has empirically documented unfairness in practical machine learning settings, identified many theoretical sources and mechanisms of unfairness, and constructed innovative fairness-aware algorithms. Researchers have developed many innovative technical solutions to these problems, yet the issue in practice remains far from solved. This Chapter highlights a simple and important point: while technical solutions to unfairness are certainly important, mitigating unfairness in practice may require tackling *economic* incen-

[1] This Chapter is based on joint work with Ben Fish [50].

tives promoting unfairness.

Most of the existing literature assumes that a fixed dataset, possibly biased, arrives in the hands of a data scientist, and solutions often revolve around clever ways to mitigate this bias. In practice, however, economic incentives may create disparities well before the data scientist enters the picture. Consider, for example, the task of speech recognition: producing accurate models may require a large amount of data, and data from speakers with accented or rarer dialects may be more costly to collect. If the market size of a minority group is small relative to the costs a firm would expend in developing accurate speech recognition software, it is likely that the group will be served with lower quality products.

In this chapter, we model the unfairness that arises when data-driven, profit-maximizing firms choose to differentially invest in data collection across groups, creating unequal error rates. In order to focus on this specific source of unfairness, we use a simple framework that elides the many other sources of bias that can seep into the machine learning pipeline. For instance, we assume that firms have unlimited budgets to purchase data at a cost from group-specific data sources of potentially infinite quantity. We also assume that both firms and users benefit from more accurate models, so that incentives are aligned. Furthermore, we assume that firms must build separate models for each group, to avoid unfairness that may come from fitting to the majority.

In order to construct our models, we borrow from the tools of learning theory and microeconomics to build simple, stylized models with crisp predictions of quantifiable unfairness. We assume each profit-maximizing firm faces a known demand curve as a function of the *worst-case* error rates for each group. Standard results from learning theory allow us to model worst-case error rates as a function of the amount of data the firm buys. We investigate three models of demand: linear demand, demand proportional to error rates, and (approximately) rational demand. For the precise description of our models and these assumptions, see Section 4.3.

We show in Section 4.4 that a profit-maximizing monopolist will choose to serve minorities (as defined by their market power) with lower quality models. Assuming linear demand, an oft-used benchmark in the economics literature, we quantify the difference in relative model quality between groups as a function of their market size, elasticity, and cost of data.

We then consider two classical remedies to the ills of monopolies: competition and regulation. Under two natural models of competition – multilinear demand (Section 4.5.1) and proportional demand (Section 4.5.2) – introducing competition does not mitigate inequality, and proportional demand even exacerbates it. Only a model in which all consumers choose the firm with (even infinitesimally) smaller error until firms reach sufficient accuracy suggests that competition will mitigate inequality (Section 4.5.3); to do so, however, this model assumes a stringent notion of rationality that may not be reflective of consumer behavior in the real world.

Given that our most plausible models suggest that competition does improve the situation, we ask whether regulation could be used to mitigate error inequality by design. In particular, in Section 4.6 we examine two simple kinds of constraints: a 'relative equality' constraint where error rates across groups must be multiplicatively close to each other, and an 'absolute equality' constraint where error rates across all groups must be sufficiently small, but may be far apart from each other. We then formally quantify the costs to profits (and when relevant, to the majority group's error rate) as a function of the threshold chosen. Finally, we conclude with takeaways, limitations, and directions for future work in Section 4.7.

4.2 Related Work

Motivation for our work comes from the many documented instances of disparity in learned model performance between groups. The existing literature has demonstrated troubling disparities in a number of domains, including incentive-aligned

domains (where both the firms and users receive benefit from more accurate models) that are the focus of this work. Wilson et al. [125] studies the performance of state-of-the-art object recognition systems, intended for applications like autonomous vehicles, and find that systems fail to recognize darker-skinned persons at much higher rates than lighter ones. Sweeney shows that search engine queries of black-associated names generated about four times the likelihood of ads for arrest records [116]. Blodgett and O'Connor show that on both complicated tasks like parsing and simple tasks like language identification, texts from speakers of African American English see vastly higher error rates [20]. Buolamwini and Gebru show that commercial facial recognition software systems misclassify race and gender among dark-skinned females at orders of magnitude higher rates than light-skinned males [25]. Mehrotra et al. [95] and Ekstrand et al. [49] identify differing satisfaction levels by age and gender in recommendation systems. The list goes on.

Researchers have engaged in many empirical and theoretical investigations to understand why these instances of unfairness occur, with the hope of developing solutions to mitigate them. Much of this work focuses on the learning algorithm itself as the source of unfairness, and attempts to incorporate fairness notions into the algorithm [82]; see e.g. Verma and Rubin [122] for a survey of fairness definitions. Training data has also been identified as source of unfairness; for example, Chen et al. identify sample size differences as a crucial source of unfairness, and decomposes induced unfairness into bias, variance, and noise [32]. Various feedback loops stemming from historical bias have also been identified as sources of unfairness [52, 51, 91]. There are a few others, including selection bias [78], using the wrong metric [100], or using a single model across multiple underlying data generating processes [84]. However, to the best of our knowledge, market forces in data investment have seen little attention as a source of unfairness. See the survey of Cowgill and Tucker [38] for an in-depth survey of perspectives on the sources of unfairness from computer science and economics.

Our models are built on insights from two extensive, and historically separate, literatures: the formalization of learning from data embodied in computational and statistical learning theory, and models of strategic interactions from the theory of industrial organization (see e.g. Tirole [119]). From learning theory, we apply fundamental bounds on sample complexity derived from the *Probably Approximately Correct* (PAC) framework (see e.g. Kearns and Vazirani [85]) to relate firms' costs to worst-case error rates; from industrial organization, we modify widely used models of demand (such as linear demand, multilinear models of imperfect substitutes [17], the Tullock contest [120], and Bertrand competition [114]) to link firms' choices to consumer behavior.

Recently, these two fields have drawn closer, as both computer scientists and economists have begun to model markets for information and data. For example, Aridor et al. [6] and Mansour et al. [93] consider the exploration-exploitation trade-offs faced by firms competing to win users in a bandits setting, while Ben-Porat and Tennenholtz formalize competition in the *prediction space* that can lead to models very different than those produced by empirical risk minimization algorithms [14, 15]. To the best of our knowledge, however, this is the first work to apply learning theory and industrial organization to explore differing incentives in the context of fairness. The work of Dong et al. [42] is the closest in form to ours, and uses a similar high-level abstraction of learning theory, as well as a proportional-error split in market share, but primarily explores questions of market concentration.

4.3 Consumer Behavior and Learning Theory

We begin by describing our framework at a high level. In our models, firms use data to create a classifier (or other machine learning model) that is then used to serve consumers. Consumers are split into non-overlapping groups, and choose a firm based on how well the firm's model is performing for their group. Firms receive

revenue based on how many consumers they attract, but must pay for the amount of the data they buy. The more data, the better their model. The firms aim to maximize their profits. In the case where there are multiple firms, the goal of each firm is to maximize their profit at equilibrium, as other firm's choices affect the number of consumers that they get, and hence their choices. Here, the firm's only (strategically relevant) choice is how much data to buy.

We start with the monopoly case, where there is only one firm. The firm chooses a number of data points M_g to buy for each group, where we write M for the vector of these choices; we write $\varepsilon_g(M_g)$ for the worst-case error the firm can guarantee for group g, and assume this error is known to consumers. The groups then respond by entering the market according to a demand function $D_g(\varepsilon_g)$, where $D_g(\varepsilon_g)$ represents the proportion of g that uses the firm's model. Each group has μ_g total people, so the firm's revenue is $\sum_{g\in\mathcal{G}} \mu_g D_g(\varepsilon_g(M_g))$. The firm also pays for the data, represented by a cost function $C(M)$.

We will discuss our choices for ε, D, and C in Sections 4.3.1 and 4.3.2. But for now, the firm's profit is just the revenue the firm makes minus the cost it spends to acquire that data, leading to the following optimization problem:

Definition 4.3.1 (The Monopolist's Problem)**.** The firm chooses M to maximize its profit $\pi(M)$, i.e.

$$\max_M \pi(M) = \max_M \sum_{g\in\mathcal{G}} \mu_g D_g(\varepsilon_g(M_g)) - C(M).$$

Because we will assume in Section 4.3.1 that ε_g is a deterministic function of M_g, we can also rewrite this optimization problem as

$$\max_\varepsilon \pi(\varepsilon) = \max_\varepsilon \sum_g \mu_g D_g(\varepsilon_g) - C(\varepsilon),$$

where ε is the vector of ε_g. We define $\pi(\varepsilon) = \sum_g \mu_g D_g(\varepsilon_g) - C(\varepsilon)$ as the total profit the monopolist makes. We will have an additively separable cost function in

g, i.e. $C(\varepsilon) = \sum_{g \in \mathcal{G}} C_g(\varepsilon_g)$, which will allow us to also refer to the per-group profit: $\pi_g(\varepsilon_g) = \sum_{g \in \mathcal{G}} \mu_g D_g(\varepsilon_g) - C_g(\varepsilon)$.

On the other hand, when there are multiple firms $\mathcal{F}$, maximizing profit is not longer just an optimization problem, because each firm's optimal choice will depend on its opponent's choice. So instead, we search for a Nash equilibrium, which is the workhorse solution concept in classical game theory. Under such a Nash equilibrium, each firm plays their best response, given all the choices of the other firms. For a more thorough background, see [59].

Extending our notation, we have the same components as in the monopolist case, except now we write M_{gi} for the number of data points the ith firm buys for group g, ε_{gi} is the error rate of the ith firm on group g, and $D_{gi}(\varepsilon_g) = D_{gi}(\varepsilon_{gi}, \varepsilon_{g,-i})$ is the demand for the ith firm from group g, given the vector $\varepsilon_g = (\varepsilon_{gi})_{i \in \mathcal{F}}$ of error rates.

Definition 4.3.2 (The Competitor's Problem)**.** Firms simultaneously announce their choices, resulting in a matrix $M = (M_{gi})$ of data points purchased. Each group in the market responds according to $\varepsilon_g(M_g)$.

Then a (pure) equilibrium under profit-maximizing firms is a set of vectors M_i^* chosen with a best response:

For all i,

$$M_i^* = \text{argmax}_{M_i} \pi_i(M_i, M_{-i}^*)$$

where for any M,

$$\pi_i(M_i, M_{-i}) = \sum_g \mu_g D_{gi}(\varepsilon_{gi}(M_{gi}), \varepsilon_{g,-i}(M_{g,-i})) - C(M_i).$$

We will only consider pure strategy Nash equilibria in this work.

Again, we can write an equivalent definition of the competitor's problem in terms of error:

$$\max_{\varepsilon_i} \pi_i(\varepsilon_i, \varepsilon_{-i}^*) = \max_{\varepsilon_i} \sum_{g \in \mathcal{G}} \mu_g D_{gi}(\varepsilon_{gi}, \varepsilon_{g,-i}^*) - C(\varepsilon_i),$$

where ε_i^* is the vector of error rates given by the associated equilibrium choice M_i^*. We also use $\pi_{gi}(\varepsilon_{gi}, \varepsilon^*_{g,-i}) = \mu_g D_{gi}(\varepsilon_{gi}, \varepsilon^*_{g,-i}) - C_g(\varepsilon_i)$ to refer to the profit i makes on group g.

Note that a firm i only enters a market in the first place if $\pi_i(\varepsilon^*) > 0$. In this work, we do not consider the case when $\pi_i(\varepsilon^*) \leq 0$, as our goal is to show that even when firms *do* enter the market for each group, market forces may *still* create a disparity between groups.

Finally, in Section 4.6, we discuss imposing regulation on a monopolist to ensure some kind of 'fairness' across groups. We consider two different kinds of constraints a regulator could impose on a firm. The first is what we refer to as *relative error equality*, which roughly corresponds to group fairness in binary classification [16] For all $g, g' \in \mathcal{G}$, we require

$$\frac{\varepsilon_g}{\varepsilon_{g'}} \leq (1 + \chi),$$

for parameter $\chi \geq 0$. On the other hand, we could ask for an *absolute error guarantee*, requiring that the error rates for both firms are low, regardless of how close to each other they are: For all $g \in \mathcal{G}$, we require instead

$$\varepsilon_g \leq \chi.$$

This roughly corresponds with maximin notions of fairness, e.g. [16, 19, 56].

We investigate what happens when a monopolist satisfies one of these two constraints. Because error is the relevant quantity from the regulator's perspective, and error and data investment are so tightly linked, we write the regulated monopolist's problem in terms of the choice of error:

Definition 4.3.3 (Regulated Monopolist's Problem)**.** The firm chooses M to maximize its profit $\pi(M)$ subject to a constraint:

$$\max_{\varepsilon} \pi(\varepsilon) = \max_{\varepsilon} \sum_g \mu_g D_g(\varepsilon_g) - C(\varepsilon) \text{ s.t. } f_r(\varepsilon) \leq 0 \ \forall r \in R,$$

where either $R = \mathcal{G} \times \mathcal{G}$ and $f_{g,g'}(\varepsilon) = \varepsilon_g - (1 + \chi)\varepsilon_{g'}$, or $R = \mathcal{G}$ and $f_g(\varepsilon) = \varepsilon_g - \chi$.

Just as is the case in binary classification, where different settings may call for different notions of fairness, which version of fairness regulator should impose will depend on the context and the ethical assumptions she maintains.

4.3.1 Data, Costs, and Learning Theory

A key component to our model is how choices in data investment drive error rates. We assume that firms build a model to provide a product to consumers, and that this model is learned from data. The firms have access to independent and identically distributed data from fixed data sources that reflect the same distribution that consumers care about.

In the PAC model of learning [85], there is a class of hypotheses H, and each hypothesis $h \in H$ has an associated risk $R(h)$, typically representing the error rate of h. For example, in binary classification, $R(h) = E_{x,y\sim\mathcal{D}}[h(x) \neq y]$, though our model will be applicable to other settings as well. With only access to data drawn from $\mathcal{D}$, rather than $\mathcal{D}$ itself, the learner cannot guarantee its risk, but *can* achieve high probability upper bounds on its risk. In the agnostic PAC setting, there is a learning algorithm that upon seeing a sample of size M, except with probability δ, returns a hypothesis h such that

$$R(h) - \min_{h' \in H} R(h') \leq K\sqrt{\frac{d_H + \log(1/\delta)}{M}},$$

where $\min_{h' \in H} R(h')$ is the Bayes error, d_H is the VC dimension of H, and K is a universal constant. (See [113] for more on PAC learning, VC dimension, and the various kinds of PAC learning.)

To model the fact that getting appropriate data can have group-dependent sources and thus costs, we assume data for each group is drawn separately from distributions $\mathcal{D}_g$. The firms choose M_g, the number of data points to draw, and will use a learning algorithm with a PAC guarantee for each data set and give to a consumer of group g the output of the corresponding hypothesis.

Achieving such bounds would not be useful to the firm unless consumers make decisions based on these bounds. Here, we assume that the consumers have no more access than the firms: they do not have access to the distribution, so they cannot make decisions based on the true group-level error rates. Given this, we assume that the consumers use firms' bound on the excess error $R(h) - \min_{h' \in H} R(h')$, which we refer to as the *worst-case excess error* rate. Of course, in reality, consumer decisions are not necessarily based on the worst-case error rate. However, given that consumers often do in practice have to make choices using relatively little information about firms, and have trouble predicting how well exactly a firm will treat them, we believe this is a natural place to start. In particular, bounds on the the excess error rates represent the minimal amount of information consumers need to make informed decisions.

Thus, we set

$$\varepsilon_{gi}(M_{gi}) = \frac{\gamma_g}{M_{gi}^{1/q}},$$

for constants $\gamma_g > 0$ and $q > 0$. For example, in agnostic PAC learning, $q = 2$ and $\gamma_g = \sqrt{K(d_H + \log(1/\delta))}$. Note that we are assuming δ is fixed ahead of time, but we allow in general for γ_g to be group-dependent. Agnostic PAC learning is far from the only type of learning to have this form; the realizable PAC setting, the multi-class setting, and many regression settings all have this form [113].

This set-up does ignore the possibility of transfer learning, i.e. using the data from D_g to help with learning for another group g'. We avoid this scenario so as to concentrate on the 'unfairness' generated via the market incentives instead of the unfairness generated, for example, when an assumption about the similarity between D_g and $D_{g'}$ fails to hold.

The choice of M_g determines not only the worst-case error rates, but the cost to the firm of generating that data, either by collecting it in the wild or buying it from another source. As mentioned above, we permit the costs to be group-dependent.

We assume the cost is additively separable and linear in M_g:

$$C_{gi}(M_{gi}) = \phi_{gi} + c_{gi}M_{gi} \text{ and } C_i(M_i) = \sum_{g \in \mathcal{G}} C_{gi}(M_{gi}),$$

for constants ϕ_{gi}, c_{gi}, where ϕ_{gi} represents the fixed cost of entering the market.

Since we can rewrite $M_{gi} = (\gamma_g/\varepsilon_{gi})^q$, this model is equivalent to first choosing a worst-case error rate ε_{gi} and then paying a cost

$$C_{gi}(\varepsilon_{gi}) = \phi_{gi} + \frac{\gamma_g}{\varepsilon_{gi}^q}$$

for each group g, where γ_g is redefined to minimize the number of constants we employ.

So now

$$M_{gi} = \frac{\gamma_g}{c_{gi}\varepsilon_{gi}^q}.$$

This version is the one we will use for the remainder of the paper. Note that the cost function is convex, which means that whenever the demand is concave, so is the profit function.

4.3.2 Models of Consumer Choice

The firm's revenue is driven by how *demand* for its product reacts to its choice of worst-case error guarantees. We consider several models of this demand, each inspired by well-studied models in microeconomics. While firms are primarily concerned only with aggregate changes in demand, rather than the decisions of individual consumers, each of our models can be founded on natural models of individual consumer behavior, and we provide such models in several cases.

In the monopoly case, we use a simple model of *linear* demand; while an idealization, linear demand is often used even in econometric estimation (see e.g. [70]). In the competitive case, there are a variety of natural demand models, each embedding different assumptions about how consumers choose between firms and how stringently they react to differing error levels. We study three models along a

spectrum of rationality: a *multilinear* demand, generalizing the monopoly case; a parameterized *proportional* demand; and an *approximately* rational demand, where consumers exclusively use the firm with lowest error (up to some tolerance).

We give the details of these models of demand in each appropriate subsection in Sections 4.4 and 4.5. Under each model, there are parameter regimes where firms choose not to invest in data collection for some groups at all ; while this may reflect some real-world scenarios, the purpose is of this Chapter is to highlight economic incentives that create inequality *even aside from such extreme scenarios*. As such, we will focus on *interior* optima or equilibria. In an interior optimum, the monopolist must make positive profit (so that it enters the market) and choose error rates strictly smaller than 1 for each group (so that it is investing in data collection for each group). Similarly, interior equilibria require that profits for both firms must be positive and each error rate strictly smaller than 1. Our theorems statements will highlight this focus.

4.4 Monopoly

We start with the case where there is one firm in the market and demand is linear:

Definition 4.4.1 (Linear Demand)**.** A *linear demand* function for each group g is given by:

$$D_g(\varepsilon_g) = \alpha_g - \beta_g \varepsilon_g,$$

where $0 < \beta \leq \alpha \leq 1$.

A linear demand curve can arise from a simple model of consumer behavior: suppose utility-maximizing consumers consider whether or not to use the product, and only use it if it is above some threshold (equivalent to being better than some 'outside option'). If these thresholds are uniformly distributed over some interval, then demand will be linearly decreasing over an interval. Strictly speaking, this

is a *piecewise* linear demand, but this does not greatly affect optimal behavior of the firm - it merely means that it will never choose outside the linearly decreasing range unless they are choosing not to invest in providing quality products at all. For simplicity of our theorem statements, we will assume that parameters are such that the firm's achievable errors are a subset of the linear portion of the demand curve, here $0 < \beta \leq \alpha \leq 1$, but in Appendix 4.9, we generalize to arbitrarily large α, β to ensure that our results still qualitatively hold.

Our main result here is the following:

Theorem 4.4.1 (Monopoly Inequality). Suppose a monopolist with learning rate q faces linear demand. Then in any interior optimum, for every pair of groups g and g', the error inequality is given by:

$$\frac{\varepsilon_g^*}{\varepsilon_{g'}^*} = \left(\frac{\mu_{g'}\beta_{g'}\gamma_g}{\mu_g\beta_g\gamma_{g'}}\right)^{\frac{1}{q+1}}.$$

Again, we focus on an interior optimum. Three factors affect the error gap between the minority and the majority: the size of the minority as a share of the total market; the marginal cost of gathering data on the minority vs. on the majority; and the elasticities of the populations with respect to the error. It is also worth noting that the fundamental nature of the learning problem, via the learning rate q, affects the magnitude of error inequality.

Theorem 4.4.1 is a consequence of the following lemma:

Lemma 4.4.2. *Suppose a monopolist with learning rate q faces linear demand. Then in any interior optimum, error levels set by the monopolist are given by:*

$$\varepsilon_g^* = \left(\frac{q\gamma_g}{\mu_g\beta_g}\right)^{\frac{1}{q+1}}.$$

Proof. Recall that

$$\pi(\varepsilon) = \sum_{g\in\mathcal{G}} \mu_g\left(\alpha_g - \beta_g\varepsilon_g\right) - \sum_{g\in\mathcal{G}}\left(\phi_g + \frac{\gamma_g}{\varepsilon_g^q}\right).$$

Now, we notice that this profit function is separable into the sum of profits from each market. Differentiating with respect to ε_g separately and setting to zero,we arrive at the first-order conditions:

$$\frac{\partial \pi}{\partial \varepsilon_g} = -\mu_g \beta_g + \frac{q\gamma_g}{\varepsilon_g^{q+1}} = 0.$$

Solving this equation yields $\varepsilon_g^* = \left(\frac{q\gamma_g}{\mu_g \beta_g}\right)^{\frac{1}{q+1}}$. This is indeed a maximum because profit is concave, so the only alternative is an exterior optimum. □

Notice that if, for all g, $\pi_g\left(\varepsilon_g^*\right) > 0$, $\pi_g\left(\varepsilon_g^*\right) > \pi_g(1)$, and $\varepsilon_g^* < 1$, then the interior optimum exists and is unique.

4.5 Competition

In this section, we show that under most reasonable models of competition, the introduction of competition does not mitigate error-inequality compared to the monopoly equilibrium, and may, in fact, increase it. Only under Bertrand-like competition, which assumes consumers are strictly rational, is inequality significantly mitigated. In particular, we show that under both the Tullock and the multi-linear models of demand, the inequality between groups as measured by the error rates does not improve relative to the monopoly case. In the case of the Tullock model, as a function of the relative size of the groups, inequality is actually worse.

4.5.1 Multilinear Demand

Next, we consider a simple generalization of linear demand to the two-firm case. This model can be interpreted as a model of competition in identical products with differing quality levels as in [9], but can also be interpreted (as well as microfounded, and used to estimate structural parameters, as in [17]) as markets for imperfect substitutes.

Definition 4.5.1 (Multilinear Demand)**.** The *multilinear linear demand* function is, for firms i and j, and for each group g,

$$D_{gi}(\varepsilon_{gi}, \varepsilon_{gj}) = \alpha_{gi} - \beta_g \varepsilon_{gi} + \lambda_g \varepsilon_{gj},$$

where $0 < \lambda_g < \beta_g \leq \alpha_{gi}$ and $\alpha_{gi} + \lambda_g \leq 1$.

We require $\beta_g > \lambda_g$ so that demand reacts more strongly to a firm's own error rates than its opponents – this ensures that if both firms increase error, total demand decreases. The other conditions on the parameters are to ensure that the demand is truly (multi)-linear, as opposed to piece-wise linear.

Note it is not the case that all consumers choose the firm with lower error, as one might expect if the products of the firms were perfect substitutes. Instead, users switch between firms depending on their error rates, and even if firms achieved perfect accuracy, the split of the total market might not be even, as captured by differing α_{gi}. This could represent brand loyalty, for example, or perhaps that firms' products are not perfectly identical.

Our main result for this case states that the gap between error rates is the same as in the monopoly case:

Theorem 4.5.1. Suppose that two firms with learning rate q compete under multilinear demand. Then in any interior equilibrium, for every pair of groups g and g', error inequality is given by

$$\frac{\varepsilon^*_{gi}}{\varepsilon^*_{g'i}} = \left(\frac{\mu_{g'}\beta_{g'}\gamma_{gi}}{\mu_g \beta_g \gamma_{g'i}}\right)^{\frac{1}{q+1}}.$$

Theorem 4.5.1 is a consequence of the fact that the firm's optimal choice does not depend on its opponent's error rates; that is they have a dominant strategy. This is formalized by the following lemma, which is enough to prove Theorem 4.5.1:

Lemma 4.5.2. *Suppose that two firms with learning rate q compete under multilinear demand. Then in any interior equilibrium, error levels are given by*

$$\varepsilon^*_{gi} = \left(\frac{q\gamma_{gi}}{\mu_g \beta_g}\right)^{\frac{1}{q+1}}.$$

Proof. This proof will be very similar to that of Lemma 4.4.2, only now, the behavior of the other firm will affect profit. Recall

$$\pi_i(\varepsilon_{gi}, \varepsilon_{gj}) = \sum_{g \in \mathcal{G}} \mu_g(\alpha_{gi} + -\beta_g \varepsilon_{gi} + \lambda_g \varepsilon_{gj}) - \sum_{g \in \mathcal{G}} \frac{\gamma_{gi}}{\varepsilon_{gi}^q}.$$

We can see that even though the behavior of the other firm will affect profit, firm i still has a dominant strategy. This is because the first-order conditions do not depend on the other firm:

$$\frac{\partial \pi_i}{\partial \varepsilon_{gi}} = -\mu_g \beta_g + \frac{q\gamma_{gi}}{\varepsilon_{gi}^{q+1}}.$$

This is the same first order condition as in the monopolist's case, with the same implication: $\varepsilon_{gi}^* = \left(\frac{q\gamma_{gi}}{\mu_g \beta_g}\right)^{\frac{1}{q+1}}$.

□

Similar to the case of the monoplist's, notice that if, for all g, i, $\pi_{gi}\left(\varepsilon_{gi}^*, \varepsilon_{gj}^*\right) > 0$, $\pi_{gi}\left(\varepsilon_{gi}^*, \varepsilon_{gj}^*\right) > \pi_{gi}(1, \varepsilon_{gj}^*)$, and $\varepsilon_{gi}^* < 1$, then an interior equilibrium exists.

4.5.2 Proportional Demand

In this section, we consider a model inspired by [42], and thus, indirectly, by the Tullock contest [120]. In particular, firms split the market proportionally to the other firms' error:

Definition 4.5.3 (Proportional Demand)**.** In a multi-firm market, we say that demand is *proportionally split* with *competition exponent* ρ if

$$D_{gi}(\varepsilon) = 1 - \frac{\varepsilon_{gi}^\rho}{\sum_j \varepsilon_{gj}^\rho} = \frac{\sum_{j \neq i} \varepsilon_{gj}^\rho}{\sum_j \varepsilon_{gj}^\rho}.$$

Here, we focus on the two-firm case, in which case we can write

$$D_{gi}(\varepsilon_{gi}, \varepsilon_{gj}) = 1 - \frac{\varepsilon_{gi}^\rho}{\varepsilon_{gi}^\rho + \varepsilon_{gj}^\rho} = \frac{\varepsilon_{gj}^\rho}{\varepsilon_{gi}^\rho + \varepsilon_{gj}^\rho}.$$

Now we can write our inequality theorem:

Theorem 4.5.2 (Inequality Under Proportional Demand). Suppose two firms with learning rate q compete under proportional demand. Then in any interior equilibrium error inequality is given by

$$\frac{\varepsilon^*_{gi}}{\varepsilon^*_{g'i}} = \left(\frac{\rho_{g'}\mu_{g'}}{\rho_g\mu_g}\right)^{\frac{1}{q}} \cdot \frac{f(\gamma_{gi},\gamma_{gj},q)}{f(\gamma_{g'i},\gamma_{g'j},q)},$$

where

$$f(\gamma_{gi},\gamma_{gj},q) = \frac{(\gamma_{gi}^q+\gamma_{gj}^q)^{\frac{2}{q}}}{\gamma_{gi}^{1-\frac{1}{q}}\gamma_{gj}^q}.$$

Recall that in the monopoly case, the exponent was $1/(q+1)$ instead of $1/q$, meaning that introducing competition under this model has actually *exacerbated* the effect of minority status on inequality. Note also that the relative inequality between two groups based on the results from a particular firm depends not only on that firm's cost structure for the two groups, but also on the opposing firm's cost structure for the two groups.

Proving Theorem 4.5.2 requires finding the equilibrium:

Lemma 4.5.4. *Suppose two firms with learning rate q face proportional demand with competition exponent ρ. In any interior equilibrium, error rates are given by:*

$$\varepsilon^*_{gi} = \left(\frac{q\gamma_{gi}}{\rho_g\mu_g}\right)^{\frac{1}{q}} \frac{(\gamma_{gi}^q+\gamma_{gj}^q)^{\frac{2}{q}}}{\gamma_{gi}^q\gamma_{gj}^q} = \left(\frac{q}{\rho_g\mu_g}\right)^{\frac{1}{q}} \frac{(\gamma_{gi}^q+\gamma_{gj}^q)^{\frac{2}{q}}}{\gamma_{gi}^{1-\frac{1}{q}}\gamma_{gj}^q}.$$

If $\varepsilon^*_{gi} < 1$ for all g and i then there exists a setting of parameters for which $(\varepsilon^*_{gi},\varepsilon^*_{gj})$ is the unique equilibrium.

For brevity, we relegate the full proof and the characterization of when these conditions hold, to Section 4.11. Below, we detail the instructive portion of the proof for the special case in which $q = \rho = 1$.

Lemma 4.5.5. *Suppose two firms with learning rate* 1 *face proportional demand with competition exponent* 1*. In any interior equilibrium, error levels are given by:*

$$\varepsilon^*_{gi} = \frac{1}{\mu_g} \cdot \frac{(\gamma_{gi} + \gamma_{gj})^2}{\gamma_{gj}}.$$

Sufficient conditions for existence are that for all g and i: $\phi_{gi} \leq \frac{\mu_g \gamma_{gj}^2}{(\gamma_{gi}+\gamma_{gj})^2}$, $\varepsilon^*_{gi} < 1$, $\frac{(\gamma_{gi}+\gamma_{gj})^2}{\mu_g \gamma_{gj}} < 1$. If, moreover, $\min_k \gamma_{gk} \geq (\max_k \gamma_{gk})^2$, then $(\varepsilon^*_{gi}, \varepsilon^*_{gj})$ is the unique equilibrium.

The conditions of Lemma 4.5.5 are stated in terms of γ_{gi}; recall, though, that γ_{gi} is not a primitive of our model, but rather the product of the per-datapoint cost and learning theory constants. These conditions thus imply conditions on these underlying constants. In the symmetric case, this asks that the per-datapoint cost c_g satisfies:

$$c_g \leq \frac{d_H + \log \frac{1}{\delta}}{12^q}.$$

which merely requires that the per-datapoint cost is not too large relative to the desired hypothesis class and success probability. In the asymmetric case, we require that firms do not face ratios of data cost to learning constant that are *too different* from each other. If either of these conditions is violated, then one or both firms may have an incentive to stop investing completely in data acquired for a group. Such non-interior equilibria can obviously lead to severe error inequality, but again, Theorem 4.5.2 demonstrates the existence of incentives to unfairness *even ruling out these extreme cases.*

Proof of Lemma 4.11.1. We can write Firm i's profit from each group as:

$$\pi_{gi}(\varepsilon_{gi}, \varepsilon_{gj}) = \mu_g \frac{\varepsilon_{gj}}{\varepsilon_{gi} + \varepsilon_{gj}} - \frac{\gamma_{gi}}{\varepsilon_{gi}} - \phi_{gj}.$$

The strategy space of the firm is to select an ε_{gi} for each group in $(0, 1]$; we search for a pure strategy Nash equilibrium. At a high level, our strategy to do so is as follows:

first, we fix the opposing firm's action ε_{gj}. Optimizing Firm i's profit gives a best-response to the fixed action ε_{gj}. An equilibrium pair must simultaneously satisfy *both* firms' first order conditions, given the other, so we obtain two simultaneous equations that yield the equilibrium relationship between the two firms' actions. Solving this yields a candidate solution. Then, we can show that the candidate solution is indeed a maximum via the concavity of the profit function. Finally, we need but check that there are no solutions at the endpoints, and we provide conditions when this is ruled out.

Now, fixing Firm j's choice ε_{gj}, the profit function π_g is just a function of ε_{gi}. Differentiating this gives:

$$\frac{\partial \pi_{gi}}{\partial \varepsilon_{gi}} = -\varepsilon_{gj}\mu_g(\varepsilon_{gi}+\varepsilon_{gj})^{-2} + \gamma_{gi}\varepsilon_{gi}^{-2}.$$

We set this equal to zero. Since satisfying this condition is required for ε_{gi} to be a best-response we can plug in ε^*_{gj}, whatever that may be, requiring:

$$\frac{\varepsilon_{gi}^2\varepsilon^*_{gj}}{\mu_g\gamma_{gi}} = (\varepsilon_{gi}+\varepsilon^*_{gj})^2, \tag{4.5.1}$$

and in particular, this must apply to the best-response ε^*_{gi}. We can apply similar logic to Firm j. Hence, for $(\varepsilon^*_{gi}, \varepsilon^*_{gj})$ to be best-responses to each other – that is, to be in (interior) equilibrium – we must have that

$$\frac{{\varepsilon^*_{gi}}^2\varepsilon^*_{gj}}{\mu_g\gamma_{gi}} = (\varepsilon^*_{gi}+\varepsilon^*_{gj})^2 = \frac{{\varepsilon^*_{gj}}^2\varepsilon^*_{gi}}{\mu_g\gamma_{gj}}. \tag{4.5.2}$$

This implies that

$$\varepsilon^*_{gj} = \varepsilon^*_{gi}\frac{\gamma_{gj}}{\gamma_{gi}}.$$

Substituting this condition back into Equation 4.5.2, we obtain that

$$\frac{{\varepsilon^*_{gi}}^3\frac{\gamma_{gj}}{\gamma_{gi}}}{\mu_g\gamma_{gi}} = (\varepsilon^*_{gi}+\varepsilon^*_{gi}\frac{\gamma_{gj}}{\gamma_{gi}})^2 \implies \varepsilon^*_{gi} = \frac{(\gamma_{gi}+\gamma_{gj})^2}{\mu_g\gamma_{gj}}.$$

Symmetric logic yields $\varepsilon^*_{gj} = \frac{(\gamma_{gi}+\gamma_{gj})^2}{\mu_g\gamma_{gi}}$.

Now, to show that this candidate solution is indeed an equilibrium, we must show that these actions are best-responses to each other. Fix $\varepsilon^*_{gj} = \frac{(\gamma_{gi}+\gamma_{gj})^2}{\mu_g \gamma_{gi}}$. Then we can view $\pi_{gi}(\varepsilon_{gi}, \varepsilon^*_{gj})$ as a continuous function on $(0, 1]$. By construction, evaluating $\frac{\partial}{\partial \varepsilon_{gi}} \pi_{gi}(\varepsilon_{gi}, \varepsilon^*_{gj})$ at ε^*_{gi} must give zero. (It is also easy to verify that this is indeed the case.) If $\pi_{i,\varepsilon^*_{gj}}(\varepsilon_{gi})$ is concave at ε^*_{gi}, then that is a local maximum of the profit function (given ε^*_{gj}).

To see that it is concave, note that

$$\frac{\partial^2}{\partial \varepsilon_{gj}^2} \pi_{gi}(\varepsilon_{gi}, \varepsilon_{gj}) = 2\varepsilon_{gj}\mu_g(\varepsilon_{gj} + \varepsilon_{gi})^{-3} - 2\frac{\gamma_{gi}}{\varepsilon_{gi}^3}.$$

Evaluating this quantity at ε^*_{gi} gives:

$$\begin{aligned}
&\frac{\partial^2}{\partial \varepsilon_{gi}^2} \pi_{gi}(\varepsilon_{gi}, \varepsilon^*_{gj})\Big|_{\varepsilon_{gi}=\varepsilon^*_{gi}} \\
&= 2\mu_g \frac{(\gamma_{gi}+\gamma_{gj})^2}{\mu_g \gamma_{gi}} \left(\frac{(\gamma_{gi}+\gamma_{gj})^2}{\mu_g \gamma_{gj}} + \frac{(\gamma_{gi}+\gamma_{gj})^2}{\mu_g \gamma_{gi}} \right)^{-3} \\
&\qquad - 2\frac{\gamma_{gi}}{((\gamma_{gi}+\gamma_{gj})^2/(\mu_g \gamma_{gi}))^3}.
\end{aligned}$$

Straightforward, if tedious, algebra lets us rewrite the right hand side and conclude that

$$\frac{\partial^2}{\partial \varepsilon_{gi}^2} \pi_{gi}(\varepsilon_{gi}, \varepsilon^*_{gj})\Big|_{\varepsilon_{gi}=\varepsilon^*_{gi}} = \frac{2\mu_g^3 \gamma_{gj}^3 \gamma_{gi}}{(\gamma_{gi}+\gamma_{gj})^6} \left[\frac{\gamma_{gi}}{\gamma_{gi}+\gamma_{gj}} - 1 \right].$$

But notice that this quantity is always negative if costs are positive; hence, ε^*_{gi} is indeed a local maximum of $\pi_{i,\varepsilon^*_{gj}}$.

To ensure that this point is a global maximum, we must compare it with the profit at the endpoint. For brevity, we defer this calculation to the Appendix in Section 4.11

Finally, note that equilibrium profits are positive if $\pi_{gi}(\varepsilon^*_{gi}, \varepsilon^*_{gj}) \geq 0$; this is true whenever

$$\frac{\mu_g \gamma_{gj}^2}{(\gamma_{gi}+\gamma_{gj})^2} \geq \phi_{gi}, \tag{4.5.3}$$

i.e. fixed costs are not extremely large. Positive profits and the fact that ε^*_{gi} globally maximizes profit given γ^*_{gj} ensures that the putative equilibrium pair forms an equilibrium.

To identify conditions in which this equilibrium is unique, we need to eliminate the only other possible equilibrium (both firms choosing $\varepsilon = 1$). Again, for brevity, we defer this calculation to the appendix.

□

Again, we pause to highlight several intuitive properties of the equilibrium. First, Firm i's choice of error for group g is *decreasing* with the market size of Group g as well as the *ferocity* of competition in Group g. These results are similar to those of Lemma 4.4.2, with a different functional form and the competition exponent of the Tullock game replacing the error elasticity of demand. It is also, intuitively, increasing in γ_{gi} and decreasing in γ_{gj}, though this is harder to see due to the functional form of f.

4.5.3 Approximately Rational Demand

Now we consider markets where consumers behave rationally. If we allow consumers to behave *fully* rationally, in the sense of always picking the firm with (even infinitesimally) smaller error, we obtain a model similar to the Bertrand model of competition [77]; accordingly, no equilibrium exists, as we show in Section 4.10.3. Hence, we instead consider a slight relaxation of the fully rational model: Suppose consumers behave rationally, *except* that they do not care about excess error up to ζ_g over the optimal error. That is, the lower firm will capture the whole market for errors that are not too small, but for $\varepsilon_{gi}, \varepsilon_{gj} \in [0, \zeta_g]$, firms again split the market.

We formally define this demand function below:

Definition 4.5.6 (Bertrand-like Tolerant Demand)**.** In a multi-firm market, we say

demand is ζ-*tolerant rational* with $\zeta > 0$ if

$$D_{gi}(\varepsilon) = \begin{cases} 1 & \min_k \varepsilon_{gk} > \zeta_g \text{ and } \varepsilon_{gi} < \min\limits_{j \neq i} \varepsilon_{gj} \\ \frac{1}{\sum_j \mathbf{1}[\varepsilon_{gj} = \min\limits_k \varepsilon_{gk}]} & \min_k \varepsilon_{gk} > \zeta_g \text{ and } \varepsilon_{gi} = \min\limits_{j \neq i} \varepsilon_{gj} \\ 0 & \min_k \varepsilon_{gk} > \zeta_g \text{ and } \varepsilon_{gi} > \min\limits_{j \neq i} \varepsilon_{gj} \\ \frac{\mathbf{1}[\varepsilon_{gi} \leq \zeta_g]}{\sum_j \mathbf{1}[\varepsilon_{gj} \leq \zeta_g]} & \min_k \varepsilon_{gk} \leq \zeta_g \end{cases}.$$

We will show that there exists a unique equilibrium here (for appropriate parameters) in which groups' error levels are determined not by their sizes, but by their optimal errors and their tolerances.

Theorem 4.5.3 (Approximately Rational Inequality). Suppose that two firms compete under ζ-tolerant demand. Then in any interior equilibrium, error inequality is given by

$$\frac{\varepsilon_g}{\varepsilon_{g'}} = \frac{\zeta_g}{\zeta_{g'}}$$

where $\zeta_g, \zeta_{g'}$ is users' tolerance threshold (assumed to be strictly positive). Moreover, if $\gamma_{gi} < \frac{\zeta_g \mu_g}{2}$ for all g, i, the unique equilibrium is interior.

In particular, Theorem 4.5.3 shows that under this approximate Bertrand-like model of competition, the dependence on group size in the error inequality is eliminated. Instead, inequality depends merely on the optimal error achievable under the hypothesis class used by firms and groups' tolerances.

Note that the conditions of Theorem 4.5.3 is just asking that

$$c_{gi} \leq \left(\frac{\mu_g \zeta_g}{2}\right)^q \frac{1}{d_{H_i} + \log \frac{1}{\delta}}.$$

As before, we can interpret this as a condition that the per-datapoint cost is not too large relative to the total market size and the learning theory constants.

Theorem 4.5.3 follows from the following lemma:

Lemma 4.5.7 (Approximate Rational Equilibrium)**.** *Suppose that two firms compete under ζ-tolerant demand, and $\gamma_{gi} < \frac{\zeta_g \mu_g}{2}$ for all g, i. Then an interior pure strategy equilibrium exists in which*

$$\varepsilon^*_{gi} = \zeta_g,$$

and this equilibrium is unique.

Proof. We posit that the profile (ζ_g, ζ_g) is an equilibrium. To see this, note that a firm deviating to some $\varepsilon > \zeta_g$ would lose its entire market share, and so would end up with negative profit. Under the conditions of the theorem, though,

$$\pi_i(\zeta_g, \zeta_{g'}) = \frac{\mu_g}{2} - \frac{\gamma_i}{\zeta_g} > 0$$

so deviating to a higher error, with negative profit, cannot be a profitable deviation. On the other hand, deviating to $\varepsilon \in [0, \zeta_g)$ would result in the same market share, but with increased costs. Hence, deviating to decreased error is also not a profitable deviation.

To see that there can be no other equilibria, notice that if both firms were setting error in $\in [0, \zeta_g)$, they would have an incentive to deviate to ζ_g; if one firm's error were in that range and the other's were above, then the firm with higher error would have an incentive to deviate to ζ_g; and finally, if both firms were above ζ_g, either firm could profitably deviate to slightly lower error. □

Unfortunately, even this relaxation of full rationality may not be a realistic model of competition in many cases; it still requires that outside of the range of $[0, \zeta_g]$, all consumers are perfectly discerning. This is unlikely to be true in practice. Without such an assumption, the conclusions of this model do not hold. Models like the proportional split and multilinear demand are more likely to capture salient market features in practice.

4.6 Regulation

In this section, we consider the perspective of a regulator with the power to require one of two kinds of error equality, and analyze the response of the monopolistic firm to each. These constraints that the regulator may impose are *relative error equality* and *absolute error equality*. We quantify the direct cost associated with imposing these constraints, in terms of increased error to the majority group under the first kind and lost profit to the monopolist in both. This serves to give a sense of the direct trade-offs involved in regulating machine-learning driven markets. We highlight, though, that there may be non-quantifiable benefits to equity across groups, and only societal deliberation can evaluate these trade-offs.

Which of these two types of regulation is preferred will depend on the context. Requiring errors across groups to all be similar – relative error equality – may not be sufficiently strong if large error is harmful regardless of another group's error rate, but also may be too strict when small absolute errors are perceived as approximately equivalent. On the other hand, absolute error equality – where we require all errors to be below a threshold – treats all small absolute errors as equivalent, but still allows a large relative gap in error rates across groups. An absolute error bound shifts the 'burden' of fairness entirely to the firm, which may be preferable from a consumer standpoint; at the same time, decreasing profits for monopolies may reduce the incentive to innovate, which may also be undesirable.

We make the following assumption for the rest of the section for ease of exposition:

Assumption 1. *There are two groups $\mathcal{G} = \{A, B\}$, there is an interior optimum $\varepsilon_A^M, \varepsilon_B^M < 1$ (i.e. the unconstrained monopoly enters the market), and B has lesser market power and higher data costs, i.e.*

$$\mu_B \beta_B \leq \mu_A \beta_A \qquad \textit{and} \qquad \gamma_B \geq \gamma_A.$$

We refer to group A as the *majority* group and B as the *minority* group. We

also define $(\varepsilon_A^M, \varepsilon_B^M)$ and $(\varepsilon_A^R, \varepsilon_B^R)$ to be the monopolist's and regulated monopolist's optimal choices, respectively.

Note that an immediate consequence of Assumption 1 and Theorem 4.4.2 is that $\varepsilon_B^M \geq \varepsilon_A^M$. Finally, we defer omitted proofs from this section to Sections 4.12 and 4.13.

4.6.1 Relative Error Equality

In this section, we imagine that a regulator requires the monopolist to achieve error rates within a bounded ratio. We will show that a monopoly responds by investing less in majority data collection and more in minority data collection than it otherwise would, resulting in worse error rates for the majority, better error rates for the minority, approximate equality between groups, and lower profits for the firm. In particular we quantify by how much error rates worsen for the majority and by how much profits are lowered for the monopolist, which we refer to as the 'price' of fairness.

We formalize the regulator's constraint as follows:

Definition 4.6.1 (Relative error equality)**.** The regulator forces the firm to achieve error guarantees of bounded ratio:

$$\frac{\varepsilon_A}{\varepsilon_B} \leq 1 + \chi \qquad \text{and} \qquad \frac{\varepsilon_B}{\varepsilon_A} \leq 1 + \chi$$

where χ is a positive constant.

As in Section 4.4, we consider a profit-maximizing monopolist. As before, each group has linear demand with market sizes μ_A and μ_B.

Now, if the regulation has 'bite' – that is, if it changes the outcome – the regulated monopolist does the minimum it can to satisfy the constraint; that is, it sets $\varepsilon_B^R = \varepsilon_A^R(1 + \chi)$. Formally:

Lemma 4.6.2 (Saturation)**.** *Suppose that the unregulated monopoly sets* $\varepsilon_B^M > \varepsilon_A^M(1+\chi)$*. Then the profit-maximizing monopoly facing the relative error constraint sets*

$$\varepsilon_B^R = \varepsilon_A^R(1+\chi).$$

The proof follows from concavity and Jensen's inequality; we provide details in 4.12.

Lemma 4.6.2 allows us to characterize the regulated monopolist's optimal choice of errors under this regulation:

Theorem 4.6.1. Suppose that the unregulated monopoly sets error $\varepsilon_B^M > (1+\chi)\varepsilon_A^M$. Then in any interior optimum, the regulated monopoly sets the errors as

$$\varepsilon_A^R = \left(q \cdot \frac{\gamma_A + \gamma_B/(1+\chi)^q}{\mu_A\beta_A + \mu_B\beta_B(1+\chi)}\right)^{\frac{1}{q+1}}$$

and $\varepsilon_B^R = (1+\chi)\varepsilon_A^R$.

Proof. By Lemma 4.6.2, $\varepsilon_B^R = (1+\chi)\varepsilon_A^R$. Thus, the profit maximization problem can be written solely as a function of ε_A:

$$\begin{aligned}\pi(\varepsilon_A) = \mu_A(\alpha_A - \beta_A\varepsilon_A) + \mu_B\left(\alpha_B - \beta_B\varepsilon_A(1+\chi)\right) \\ - \left(\phi_A + \frac{\gamma_A}{\varepsilon_A^q}\right) - \left(\phi_B + \frac{\gamma_B}{\varepsilon_A^q(1+\chi)^q}\right).\end{aligned}$$

Then, the first order condition is

$$\mu_A\beta_A + \mu_B\beta_B(1+\chi) = \frac{q\left(\gamma_A + \gamma_B/\left(1+\chi\right)^q\right)}{\varepsilon_A^{q+1}},$$

and hence we must have that

$$\varepsilon_A^R = \left(\frac{q\left(\gamma_A + \gamma_B/(1+\chi)^q\right)}{\mu_A\beta_A + \mu_B\beta_B(1+\chi)}\right)^{\frac{1}{q+1}}.$$

Concavity guarantees that this is a global optimum. □

These together provide insight into to what the regulation is doing. The monopolist's problem can be written as:

$$\max_{\varepsilon} \pi(\varepsilon) = \max_{\varepsilon} \mu_A\alpha_A + \mu_B\alpha_B - (\mu_A\beta_A + \mu_B\beta_B(1+\chi))\varepsilon \\ - (\phi_A + \phi_B) - \frac{1}{\varepsilon^q}(\gamma_A + \gamma_B/(1+\chi)^q).$$

This is equivalent to facing a *single* population of with demand function $\mu_A\alpha_A + \mu_B\alpha_B - (\mu_A\beta_A + \mu_B\beta_B(1+\chi))\varepsilon$, fixed cost $\phi_A + \phi_B$, and marginal cost $(\gamma_A + \gamma_B/(1+\chi)^q)$. We later use this interpretation to quickly calculate the constrained monopolist's profits.

One might worry that imposing fairness requires making both groups worse off in an absolute sense. It turns out that this is not the case; if the regulation has bite, then it necessarily increases the error of the majority group, *and necessarily decreases* the error of the minority group. That is, equality comes at a price for the majority group, but *does not* require a Pareto deterioration.

Our first result is that the monopolist will respond to regulation by increasing majority error rates.

Corollary 4.

$$\varepsilon_A^R \geq \varepsilon_A^M \qquad \textit{and} \qquad \varepsilon_B^R \leq \varepsilon_B^M.$$

At this point, members of the majority group may be concerned because their error rate increases. We refer to the gap between their error rates under the constrained and unconstrained monopolies as a *price of fairness* for this reason, even though imposing this constraint may be on the whole desirable from a societal perspective:

$$\text{PoF}_{1+\chi} = \frac{\varepsilon_A^R}{\varepsilon_A^M}.$$

Fortunately, we can show this price is relatively small:

Corollary 5 (Price of Fairness Upper Bound)**.**

$$PoF_{1+\chi} \leq \left(1 + \frac{\gamma_B}{\gamma_A} \cdot \frac{1}{(1+\chi)^q}\right)^{\frac{1}{q+1}}.$$

Unsurprisingly, this bound is increasing in the ratio of minority cost to majority cost and decreasing in the leniency of the regulator. Also unsurprisingly, decreasing the ratio $\frac{\mu_B}{\mu_A}$ or $\frac{\beta_B}{\beta_A}$ and increasing the ratio $\frac{\gamma_B}{\gamma_A}$ all increase the price of fairness for the majority.

If regulation changes the monopolist's behavior, it must weakly decrease profits. This loss is quantifiable as another price of fairness:

Definition 4.6.3 (Monopolist Price of Fairness, Relative Error)**.** We define the price of fairness as the ratio between the unconstrained monopoly profit and constrained monopoly profit under the relative error constraint, i.e.

$$\text{MonPoF}_{1+\chi} = \frac{\pi\left(\varepsilon_A^M, \varepsilon_B^M\right)}{\pi\left(\varepsilon_A^R, \varepsilon_B^R\right)} = \frac{\pi(\varepsilon_A^M, \varepsilon_B^M)}{\max_{\varepsilon_A, \varepsilon_B : \frac{\varepsilon_A}{\varepsilon_B} \leq 1+\chi, \frac{\varepsilon_B}{\varepsilon_A} \leq 1+\chi} \pi(\varepsilon_A, \varepsilon_B)}.$$

We can write down this price of fairness as a function of the parameters of the model:

Theorem 4.6.2. The Monopolist's price of fairness is given by

$$\text{MonPoF}_{1+\chi} = \frac{\mu_A\alpha_A + \mu_B\alpha_B - Q(\mu_A\beta_A)^{\frac{q}{q+1}}\gamma_A^{\frac{1}{q+1}} - Q(\mu_B\beta_B)^{\frac{q}{q+1}}\gamma_B^{\frac{1}{q+1}}}{\mu_A\alpha_A + \mu_B\alpha_B - Q(\mu_A\beta_A + \mu_B\beta_B(1+\chi))^{\frac{q}{q+1}}(\gamma_A + \gamma_B\frac{1}{(1+\chi)^q})^{\frac{1}{q+1}}},$$

where $Q = q^{\frac{1}{q+1}} + \frac{1}{q^{q/(q+1)}}$.

Proof. The optimal solution to the monopolist's problem with parameters $\mu_g, \alpha_g, \beta_g, \gamma_g$ for g in $\mathcal{G} = \{A, B\}$ is the following:

$$\pi^*(\varepsilon^*) = \sum_{g \in \{A,B\}} \mu_g\alpha_g - (\mu_g\beta_g)^{\frac{q}{q+1}}\gamma_g^{\frac{1}{q+1}}Q.$$

(See Appendix 4.12.) Using this form and plugging in the market parameters, we obtain the optimal profit of the unconstrained monopolist for the numerator. The denominator is derived using the interpretation of the constrained monopolist's problem as optimizing its profits against a single market with parameters modified by regulation, and plugging these parameters into the same form. □

Theorem 4.6.2 provides a quantitative *price of fairness* in terms of monopoly profits. However, it is somewhat unwieldy; Proposition 13 provides some clarity on the limiting behavior of this price of fairness as a function of the minority group's size in absolute terms.

Claim 13 (MonPoF Limit - Relative Error Inequality)**.** *Let $\mu_B/\mu_A = r$ for constant ratio r. Then*

$$\lim_{\mu_B \to \infty} MonPoF_{1+\chi} = 1.$$

On the other hand, for constant μ_A,

$$\lim_{\mu_B \to 0} MonPoF_{1+\chi} = \frac{1 - (Q/q)\varepsilon_A^M}{1 - (Q/q)\varepsilon_A^M \left[1 + \frac{\gamma_B}{\gamma_A} \frac{1}{(1+\chi)^{\frac{1}{q+1}}}\right]^{\frac{1}{q+1}}}$$

where Q is as above.

4.6.2 Absolute Error Equality

In this section, we suppose instead that the regulator imposes an absolute upper bound on error rates for each group. We show that the monopolist responds by purchasing just enough data to meet the constraint using the profits from the majority to subsidize the minority. In this case, minority error rates can be improved without increasing error for the majority; the regulator can even improve error rates for the majority as well, up to a point. We characterize the price of fairness for the monopolist and the minimum error the regulator can guarantee. We formalize this constraint as follows:

Definition 4.6.4 (Absolute error equality)**.** For $\chi < 1$, the regulator forces the firm to achieve error of:

$$\varepsilon_A \leq \chi \qquad \text{and} \qquad \varepsilon_B \leq \chi.$$

We have another saturation lemma for this kind of constraint too: either the unconstrained error was already less than χ, or the profit maximizing error subject to regulation is exactly χ. Formally:

Lemma 4.6.5 (Saturation). *$\forall g \in \{A, B\}$, if $\varepsilon_g^R \neq \varepsilon_g^M$ then $\varepsilon_g^R = \chi$.*

Lemma 4.6.5 lets us reason very simply about the behavior of the regulated monopolist: for any group in which imposing regulation requires the firm to improve error rates, the firm will use up the entirety of this 'error budget.' Profit will decrease, of course, because imposing constraints can only decrease its objective. In this scenario, if the firm enters the market at all, it must enter the market for both groups so as to achieve the constrained error rates. A regulator then has to choose χ so as to still induce the firm to enter the market at all if they want to ensure the constrained error rate for the minority group. Of course, a regulator may also wish to choose the smallest such error rate, which we refer to as the *minimum achievable error*. Lemma 4.6.5 let us characterize the minimum achievable error:

Claim 14. *Let χ_0 be the smallest $\chi \in [0, 1]$ which solves*

$$K_1\chi^{q+1} + K_2\chi^q - K_3 = 0, \tag{4.6.1}$$

where $K_1 = -(\mu_A\beta_A + \mu_B\beta_B)$, $K_2 = \mu_A\alpha_A + \mu_B\alpha_B - \phi_A - \phi_B$, and $K_3 = \gamma_A + \gamma_B$. χ_0 exists and is the minimum achievable error, i.e. the minimum $\chi \in [0, 1]$ for which the monopolist still enters the market.

Equation 4.6.1 can be solved via the quadratic or cubic formulae in the realizable and agnostic cases, respectively, and learning rates in between can be accommodated numerically. This leads us to the monopoly's optimal error rates as a function of χ:

Theorem 4.6.3 (Absolute Outcomes). Outcomes fall into one of the following possibilities:

1. If $\chi \geq \varepsilon_B^M$ then $(\varepsilon_B^R, \varepsilon_B^R) = (\varepsilon_A^M, \varepsilon_B^M)$.

2. If $\varepsilon_A^M \leq \chi < \varepsilon_B^M$ then $(\varepsilon_B^R, \varepsilon_B^R) = (\varepsilon_A^M, \chi)$.

3. If $\chi_0 < \chi < \varepsilon_A^M$ then $(\varepsilon_B^R, \varepsilon_B^R) = (\chi, \chi)$.

4. If $\chi < \chi_0$ then the firm exits the market.

Proof. Case 1 is trivial. Case 2 and 3 follow from Lemma 4.6.5. Case 4 follows by the definition of χ_0. □

Theorem 4.6.3 contrasts starkly with Theorem 4: as long as the constraint is not so strict the monopolist exits the market, outcomes either improve for the minority and remain just as good for the majority, or improve for *both* groups. In other words, this style of regulation *does not impose a price of fairness on the majority.* Note that unless $\varepsilon_0 < \chi < \varepsilon_A^M$, the regulator is *not* guaranteeing relative equality. Which type of equality is preferable will depend on the context. Of course, this regulation *does* still impact profit:

Definition 4.6.6 (Monopolist Price of Fairness)**.** We define the monopolist's *price of fairness* under absolute error constraints as:

$$\text{MonPoF}_\chi = \frac{\pi\left(\varepsilon_A^M, \varepsilon_B^M\right)}{\pi\left(\varepsilon_A^R, \varepsilon_B^R\right)} = \frac{\pi(\varepsilon_A^M, \varepsilon_B^M)}{\max_{\varepsilon_A, \varepsilon_B : \varepsilon_A \leq \chi, \varepsilon_B \leq \chi} \pi(\varepsilon_A, \varepsilon_B)}.$$

Notice that given the market parameters, Theorem 4.6.3 allows the regulator to evaluate the monopolist's price of fairness for each potential choice of error threshold via straightforward calculation. Proposition 15 characterizes the limiting behavior of the monopolist's price of fairness as a function of absolute size of the minority group under absolute error guarantees, and these are qualitatively similar to limiting behavior under relative error guarantees.

Claim 15 (MonPoF Limit - Absolute Error Guarantees)**.** *For* fixed χ*, and for* $\mu_B \to \infty$ *at a constant ratio* $\mu_A/\mu_B = r$*:*

$$\lim_{\mu_B \to \infty} \mathit{MonPoF}_\chi = 1.$$

On the other hand, let χ_0 be the minimal achievable error when $\mu_B = 0$ (i.e. when the firm faces group A alone). Then if $\chi > \chi_0$, then $MonPoF_\chi$ converges to a parameter-specific constant as $\mu_B \to 0$.

4.7 Discussion

In this work, we identify economic incentives leading to unfairness in data-driven markets. At a high level, we show that monopolists are incentivized to invest less in minority groups (as measured by market size, elasticity, and data costs) because they are less profitable; that competition does not mitigate this incentive towards inequality, under reasonable models; and that judicious regulation *can* improve outcomes, potentially at a cost in terms of profits or, depending on the regulation, error rates for the majority group.

We view this Chapter as highlighting an important and understudied point of view, but certainly not as the last word. We made many choices that situate our models in particular contexts; for example, the assumption that firms and users benefit from improved accuracy does not capture many settings that currently are or will soon be urgent domains of adjudicating fairness concerns - machine learning in loans, insurance, and facial recognition systems are obvious cases, but the potential, and consequent scope for unfairness, is vast. We hope that future work will further clarify the possibility - and perhaps necessity- of leveraging policy tools in addition to algorithmic solutions to combat unfairness in machine learning.

4.8 Appendix

4.9 Piece-wise linear demand

In this section, we consider the possibility of a *piece-wise* linear demand function. Such a demand has the same spirit of the linear demand function, in that market share declines linearly with worst-case error rate, but allows for a more general parameter range. In particular, a firm with piece-wise linear demand may capture less than the full market (but, logically, not more) with perfect accuracy, and may lose the entire market even at relatively high accuracy. Imposing a cap and floor on a linear demand function whose parameters fall outside the restricted range described in Section 4.4 allows us to accomplish this.

We formally write piece-wise linear demand as follows:

$$D_g(\varepsilon_g) = \begin{cases} 0 & \text{if } \varepsilon_g \geq \alpha_g/\beta_g \\ \alpha_g - \beta_g \varepsilon_g & \text{if } \frac{\alpha_g - 1}{\beta_g} \leq \varepsilon_g \leq \alpha_g/\beta_g \\ 1 & \text{if } \varepsilon_g \leq \frac{\alpha_g - 1}{\beta_g}, \end{cases}$$

where $\alpha_g, \beta_g > 0$, and $\frac{\alpha_g - 1}{\beta_g} < 1$.

Finding the optimal choice of the monopolist under this demand requires slightly more care than linear demand but is substantively similar. We provide an outline below.

Lemma 4.9.1. *For convenience, let*

$$\tilde{\varepsilon}_g = \max\left\{\min\left\{\left(\frac{q\gamma_g}{\mu_g \beta_g}\right)^{\frac{1}{q+1}}, \frac{\alpha_g}{\beta_g}, 1\right\}, \frac{\alpha_g - 1}{\beta_g}\right\}.$$

A monopolist, under linear demand enters the market for group g if and only if

$$\pi_g\left(\tilde{\varepsilon}_g\right) > 0,$$

and if they do, the equilibrium error rate ε_g^* is:

$$\varepsilon_g^* = \tilde{\varepsilon}_g \qquad .$$

Proof outline. Since the profit is additively separable over g, we consider each π_g separately. For $\varepsilon_g \geq \alpha/\beta$, note profit is always negative. And for $\varepsilon_g \leq \frac{\alpha_g - 1}{\beta_g}$, demand is increasing as ε_g increases, which can be seen by checking the derivative. Then if profits are positive, $\frac{\alpha_g - 1}{\beta_g} \leq \varepsilon_g^* \leq \alpha_g/\beta_g$. Thus either ε^* is one of those end points, or ε^* satisfies the first order condition $\varepsilon_g^* = \varepsilon_g : \frac{\partial \pi}{\partial \varepsilon_g} \mid_{\varepsilon_g} = 0$, as in Lemma 4.4.2, and thus $\varepsilon_g^* = \tilde{\varepsilon}_g$.

Moreover, if the maximum profit is positive, it must be attained with $\varepsilon_g^* \leq \alpha_g/\beta_g$, so it must be the case that the profit obtained at $\tilde{\varepsilon}_g$ is positive, and vice versa.

□

4.10 Consumer Models

In this section, we show how natural models of consumer behavior give rise to the demand functions we assumed for our analysis.

4.10.1 Linear Demand

First, consider the following interaction between one firm and a representative user: The firm sets its error levels; the user uses the service if they will receive an accurate answer with probability higher than some threshold corresponding to their outside option (i.e. the payoff they would get if they decide not to use the service). While the user knows her outside option, the firm does not; a standard approach is to assume the firm makes decisions as if the user's outside option were drawn from a *distribution*. If this distribution is *uniform* over some interval, then there is a linear relationship between choice of error and probability (from the firm's perspective) of the user choosing to use the service (and thus the firm's expected revenue). If the firm interacts with many users, and these threshold are uniform throughout the population, then this representative interaction captures the aggregate interaction the firm faces.

We formalize the interaction as follows: A firm provides a service to a *user* wishing to answer some query. If the response is accurate, the user receives a payoff of 1; otherwise, 0. The firm's worst-case error rate ε is known to the user, and the user chooses whether or not to use the firm's service based on their expected utility under the worst-case error. The user has some parameter, τ, describing their payoff from choosing not to use the service. This parameter is drawn from the uniform distribution over $[\underline{\tau}, \bar{\tau}]$, that describes their outside option distribution.

To see the correspondence between this model and linear demand, we claim that any linear demand function $D(\varepsilon) := \alpha - \beta\varepsilon$ can be mapped to the probability that a user uses the service under some particular choice $[\underline{\tau}, \bar{\tau}]$. Formally:

Claim 16. *For any linear demand function $D(\varepsilon) = \alpha - \beta\varepsilon$, there exists a uniform outside option model with choice $\underline{\tau} = 1 - \frac{\alpha}{\beta}$, $\bar{\tau} = 1 + \frac{1-\alpha}{\beta}$ that justifies it.*

Proof. To see this, first note that the user will use the service if and only if the expected payoff is less than his outside option. Since the user receives a payoff of 1 if the service answers correctly and 0 if it answers incorrectly, the expected payoff is merely $1 * \Pr[\text{correct}] + 0 * \Pr[\text{incorrect}] = 1 - \varepsilon$. Hence, the user will use the service if and only if $1 - \varepsilon \geq \tau$. Now, since the user's outside option is, from the Firm's perspective, a uniform random variable, the probability that the user will use the service, as a function of ε, can be written as:

$$\begin{aligned}
\Pr[\text{user uses}](\varepsilon) &= \Pr_{\tau \sim U[\alpha,\beta]}[\varepsilon < 1 - \tau] \\
&= \Pr_{\tau \sim U[\alpha,\beta]}[\tau < 1 - \varepsilon] \\
&= \frac{1 - \varepsilon - \underline{\tau}}{\bar{\tau} - \underline{\tau}} = \frac{1 - \underline{\tau}}{\bar{\tau} - \underline{\tau}} - \frac{\varepsilon}{\bar{\tau} - \underline{\tau}}.
\end{aligned}$$

Letting $\alpha = \frac{1-\underline{\tau}}{\bar{\tau}-\underline{\tau}}$, $\beta = \frac{1}{\bar{\tau}-\underline{\tau}}$ and solving for $\bar{\tau}$ and $\underline{\tau}$ yields the claim. □

Notice that the truth of the claim is a matter of algebra and holds even beyond sensible choices for α and β. That is, choosing $\alpha > 1$ would still map to a plausible instance of linear demand, but $\alpha > 1$ would not be sensible as the intercept for

a linear probability model. Finally, notice that the simple case of $\alpha = 1, \beta = 1$ corresponds to the uniform random variable over $[0, 1]$.

4.10.2 Proportional Split

Consider the following Markov chain representing plausible user behavior in the presence of competition: at any time t, a user who is currently using Firm i stays with Firm i into time $t + 1$ if the firm does not make a mistake; otherwise, the user switches to Firm j with probability α and leaves the market with probability $1 - \alpha$. A user outside the market re-enters it with probability β, and then chooses uniformly from the firms.

The steady state distribution of this Markov chain solves the following equations:

$$\begin{aligned}
\mu_1 &= (1 - \varepsilon_1)\mu_1 + \alpha\varepsilon_2\mu_2 + \frac{\beta}{2}\mu_3 \\
\mu_2 &= \alpha\varepsilon_1\mu_1 + (1 - \varepsilon_2)\mu_2 + \frac{\beta}{2}\mu_3 \\
\mu_3 &= (1 - \alpha)\varepsilon_1\mu_1 + (1 - \alpha)\varepsilon_2\mu_2 + (1 - \beta)\mu_3.
\end{aligned}$$

Viewing the firm's market share as the proportion of times the user chooses the firm over a long enough horizon (or over many enough consumers) yields a correspondence between the market share and the stationary distribution. The form of this correspondence follows from the following lemma:

Lemma 4.10.1. *Firm i's market share under this Markov process is given by*

$$\mu_i = \frac{\varepsilon_j}{\varepsilon_i + \varepsilon_j + \tau\varepsilon_i\varepsilon_j}.$$

where $\tau = 2\frac{1-\alpha}{\beta}$.

Proof. We first the original three equations characterizing the steady state distri-

bution as:

$$\mu_1 = \frac{1}{\varepsilon_1}\left[\alpha\varepsilon_2\mu_2 + \frac{\beta}{2}\mu_3\right]$$
$$\mu_2 = \frac{1}{\varepsilon_2}\left[\alpha\varepsilon_1\mu_1 + \frac{\beta}{2}\mu_3\right]$$
$$\mu_3 = \frac{1-\alpha}{\beta}\left[\varepsilon_1\mu_1 + \varepsilon_2\mu_2\right].$$

Solving the first two equations for μ_3 and setting equal to each other requires that

$$\frac{2\varepsilon_1}{\beta}[\mu_1 - \alpha\frac{\varepsilon_2}{\varepsilon_2}\mu_2] = \frac{2\varepsilon_2}{\beta}[\mu_2 - \alpha\frac{\varepsilon_2}{\varepsilon_2}\mu_1] \implies \varepsilon_1\mu_1 = \varepsilon_2\mu_2.$$

Substituting this into the first rewritten equation for μ_3 gives that

$$\mu_3 = \frac{1-\alpha}{\beta}2\varepsilon_1\mu_1.$$

Finally applying the constraint that $\mu_1 + \mu_2 + \mu_3 = 1$ implies that

$$\mu_1 + \frac{\varepsilon_1}{\varepsilon_2}\mu_1 + 2\varepsilon_1\frac{1-\alpha}{\beta}\mu_1 = 1 \implies \mu_1 = \frac{1}{1 + \frac{\varepsilon_1}{\varepsilon_2} + 2\frac{1-\alpha}{\beta}\varepsilon_1}.$$

Now, we can reparameterize $2\frac{1-\alpha}{\beta}$ as τ, and apply the symmetric logic to the other firm to obtain the general result:

$$\mu_i = \frac{\varepsilon_j}{\varepsilon_i + \varepsilon_j + \tau\varepsilon_i\varepsilon_j}.$$

Thus, viewing the market share of Firm i as its share of the stationary distribution gives the result claimed. □

Notice that the case of $\alpha = 1$ recovers the case in which firms split the complete market, and we can again consider integral competition exponents as requiring ρ mistakes in a row before switching. In this Chapter, we only consider the case in which $\alpha = 1$.

4.10.3 Fully Rational Demand

The Bertrand model of competition considers firms competing on price with *fully rational* consumers. These consumers will always pick the firm with (even infinitesimally) lower price. It is known that a Nash equilibrium exists when firms have identical constant marginal costs in quantity and can produce an unlimited quantity. In that case, firms set equilibrium price equal to marginal cost (that is, the lowest price that firms could charge without losing money). We modify the Bertrand model to apply to our setting. Firms do not set prices in our model; instead, they change error rates. This is not a perfect analogy – changing error rates is itself costly – but captures the spirit of the Bertrand model. However, as we show in this section, equilibrium need not exist in the fully rational model (just as a pure-strategy equilibrium need not exist in canonical Bertrand competition when firms face non-constant marginal costs).

Informally, we say that demand is *fully rational*, or Bertrand-like, if firms with the minimum error capture the entire market (with ties broken by splitting the market equally).

Definition 4.10.2 (Fully Rational Demand)**.** In a multi-firm market, we say that demand is *fully rational* if

$$D_{gi}(\varepsilon) = \begin{cases} 1 & \varepsilon_{gi} < \min\limits_{j \neq i} \varepsilon_{gj} \\ \frac{1}{\sum_j \mathbf{1}[\varepsilon_{gj} = \min\limits_k \varepsilon_{gk}]} & \varepsilon_{gi} = \min\limits_{j \neq i} \varepsilon_{gj} \\ 0 & \varepsilon_{gi} > \min\limits_{j \neq i} \varepsilon_{gj} \end{cases}.$$

A proposition we will show is that there is no equilibrium in pure strategies when considering this fully-rational demand.

Claim 17. *The game induced by* fully rational demands *as described in Definition 4.10.2 has no equilibrium in pure strategies whenever* $c_{gi} < \frac{\mu_g}{2}$ $\forall i$ *for some group* g*.*

Proof. Suppose there existed such an equilibrium. Consider a single group and let $(\varepsilon^*_{gi}, \varepsilon^*_{gj})$ be the putative equilibrium error choices. Note that these correspond to equilibrium choices of data (M^*_{gi}, M^*_{gj}). We claim that a profitable deviation will exist regardless of what these choices are. There are two cases: in the first, firms have different errors, while in the second, firms have the same error. If firms have different errors, without loss of generality suppose that $\varepsilon^*_{gi} < \varepsilon^*_{gj}$. Then Firm i receives $\mu_g - \gamma_{gi}/\varepsilon^q_{gi} - \phi_{gi}$, while Firm j attains zero revenue. But notice that Firm i can unilaterally deviate to $\varepsilon' \in (\varepsilon^*_{gi}, \varepsilon^*_{gj})$ and capture the full market while paying less, thus improving profits. Hence, we cannot have an equilibrium when firms are choosing different error rates. On the other hand, suppose firms are choosing the same error rates, that is, $\varepsilon^*_{gi} = \varepsilon^*_{gj}$. Now, we can link ε^*_{gi} to M^*_{gi} via $\varepsilon^*_{gi} = \frac{(d_H + \log\frac{1}{\delta})}{(M^*_{gi})^{\frac{1}{q}}}$. In this case, each firm is earning $\frac{\mu_g}{2} - c_g M^*_{gk} - \phi_{gk}$. Consider Firm i buying an additional data point, i.e. $M'_{gi} = M^*_{gi} + 1$.Then because worst-case error guarantees are strictly decreasing in the number of datapoints purchased, we must have that $\varepsilon'_{gi} < \varepsilon^*_{gi}$, and thus the firm deviating to M'_{gi} would capture the whole market at a cost of $c_{gi}(M^*_{gi} + 1)$. This deviation will be profitable if

$$\mu_g - c_{gi}(M^*_{gi} + 1) > \frac{\mu_g}{2} - c_{gi} M^*_{gi} \iff \frac{\mu_g}{2} > c_{gi}.$$

Thus if $c_{gi} < \frac{\mu_g}{2}\ \forall i$, $(\varepsilon^*_{gi}, \varepsilon^*_{gj})$ cannot be an equilibrium. □

A natural way to relax full rationality is to allow consumers to be rational *up to a point*. That is, above some threshold ξ, they can perfectly discriminate between error rates, and always will choose the firm with (even infinitesimally) smaller error. But below ξ, increasing accuracy does not materially improve their utility of the project, and rather than attempt to ferret out small differences, they pick randomly among firms with error below ξ. This leads to our ξ-*tolerant rational demand* as discussed in Section 4.5.3.

4.11 Omitted Proofs from Section 5

Remainder of Proof of Lemma 4.5.5. The profit of playing ε^*_{gi} given that j chooses ε^*_{gj} is

$$\pi_{gi}(\varepsilon^*_{gi}, \varepsilon^*_{gj}) = \frac{\mu_g \gamma^2_{gj}}{(\gamma_{gi} + \gamma_{gj})^2} - \phi_{gi}.$$

On the other hand, if the firm chooses $\varepsilon_{gi} = 1$, its profit can be upper bounded as:

$$\pi_{gi}(1, \varepsilon^*_{gj}) \leq \mu_g \frac{(\gamma_{gi} + \gamma_{gj})^2}{\gamma_{gi} + (\gamma_{gi} + \gamma_{gj})^2} - \phi_{gi}.$$

Hence, their difference is at least:

$$\pi_{gi}(\varepsilon^*_{gi}, \varepsilon^*_{gj}) - \pi_{gi}(1, \varepsilon^*_{gj}) =\geq \mu_g \left[\frac{\gamma^2_{gj}}{(\gamma_{gi} + \gamma_{gj})^2} - \frac{(\gamma_{gi} + \gamma_{gj})^2}{\gamma_{gi} + (\gamma_{gj} + \gamma_{gi})^2} \right].$$

Thus, a sufficient condition that $\pi_{gi}(\varepsilon^*_{gi}, \varepsilon^*_{gj}) \geq \pi_{gi}(1, \varepsilon^*_{gj})$ is:

$$\frac{(\gamma_{gi} + \gamma_{gj})^2}{\gamma_{gi} + (\gamma_{gi} + \gamma_{gj})^2} < \frac{\gamma^2_{gj}}{(\gamma_{gi} + \gamma_{gj})^2}.$$

This is true if and only if:

$$(\gamma_{gi} + \gamma_{gj})^4 < \gamma^2_{gj} \left[\gamma_{gi} + (\gamma_{gi} + \gamma_{gj})^2 \right]. \tag{4.11.1}$$

On the other hand, to ensure that ε^*_{gj} is a best-response to ε^*_{gi}, we carry out the symmetric logic for Firm j. This will require that

$$(\gamma_{gi} + \gamma_{gj})^4 < \gamma^2_{gi} [\gamma_{gj} + (\gamma_{gi} + \gamma_{gj})^2]. \tag{4.11.2}$$

Both inequalities must be satisfied if our purported equilibrium is to be truly an equilibrium. Characterizing possible simultaneous solutions to Inequalities 4.11.1 and 4.11.2 is tedious, so instead we note that it suffices to ensure $\min_k \gamma_{gk} \geq 12(\max_k \gamma_{gk})^2$; to avoid encumbering the current argument, we defer the proof of

this fact to Appendix 4.11. These are not the *only* solutions conditions that satisfy Inequality 4.11.1, but they are sufficient conditions convenient to write down.

Finally, note that equilibrium profits are positive if $\pi_{gi}(\varepsilon^*_{gi}, \varepsilon^*_{gj}) \geq 0$; this is true whenever

$$\frac{\mu_g \gamma_{gj}^2}{(\gamma_{gi} + \gamma_{gj})^2} \geq \phi_{gi}, \tag{4.11.3}$$

i.e. fixed costs are not extremely large.

Thus, $(\varepsilon^*_{gi}, \varepsilon^*_{gj})$ satisfy these three conditions- Inequality, 4.11.1, Inequality 4.11.2, and Inequality 4.11.3 - and $\varepsilon^*_{gi} < 1$; hence $(\varepsilon^*_{gi}, \varepsilon^*_{gj})$ is truly an interior equilibrium. □

Technical Lemma for Simple Tullock Case

We now supply the missing algebra from Lemma 4.11.1:

Lemma 4.11.1 (Technical Lemma)**.** *The inequalities:*

$$(\gamma_{gi} + \gamma_{gj})^4 < \gamma_{gj}^2 \left[\gamma_{gi} + (\gamma_{gi} + \gamma_{gj})^2\right],$$
$$(\gamma_{gi} + \gamma_{gj})^4 < \gamma_{gi}^2 \left[\gamma_{gj} + (\gamma_{gi} + \gamma_{gj})^2\right]$$

(inequalities 4.11.1 and 4.11.2) will be satisfied if $\min_k \gamma_{gk} \geq 12(\max_k \gamma_{gk})^2$*. In the symmetric case, then* $\gamma_g \leq \frac{1}{12}$*.*

Proof. The set we are interested in is the intersection of two solution sets to polynomial equations, and is hard to characterize precisely; however, we can give sufficient conditions on γ_{gi}, γ_{gj} so that both inequalities are simultaneously satisfied.

We begin with the *symmetric* case, where $\gamma_{gi} = \gamma_{gj} = \gamma$, as it is easy to see: this is asking that

$$2^4\gamma^4 < \gamma^3 + 2^2\gamma^2\gamma^2 \iff 12\gamma^4 < \gamma^3 \iff \gamma < \frac{1}{12}. \tag{4.11.4}$$

If, instead, $\gamma_{gi} \neq \gamma_{gj}$, then we need to examine the algebra more carefully. We claim that the if $\max\limits_k \gamma_k < \frac{1}{16}$ and $\min\limits_k \gamma_{gk} > (\max\limits_k \gamma_{gk})^2$ will suffice.

To see this, note that Inequality 4.11.1 expanded out is:

$$\begin{aligned}
0 &< \gamma_{gj}^2(\gamma_{gi} + (\gamma_{gj} + \gamma_{gi})^2) - (\gamma_{gi} + \gamma_{gj})^4 \\
&= \gamma_{gj}^2\gamma_{gi} + \gamma_{gj}^2(\gamma_{gi}^2 + \gamma_{gj}^2 + 2\gamma_{gi}\gamma_{gj}) \\
&\qquad - (\gamma_{gi}^4 + 4\gamma_{gi}^3\gamma_{gj} + 6\gamma_{gi}^2\gamma_{gj}^2 + 4\gamma_{gi}\gamma_{gj}^3 + \gamma_{gj}^4) \\
&= \gamma_{gj}^2\gamma_{gi} + \gamma_{gj}^2\gamma_{gi}^2 + \gamma_{gj}^4 + 2\gamma_{gi}\gamma_{gj}^3 \\
&\qquad - \gamma_{gi}^4 - 4\gamma_{gj}^3\gamma_{gj} - 6\gamma_{gi}^2\gamma_{gj}^2 - 4\gamma_{gi}\gamma_{gj}^3 - \gamma_{gj}^4 \\
&= \gamma_{gj}^2\gamma_{gi} - 5\gamma_{gj}^2\gamma_{gi}^2 - 2\gamma_{gi}\gamma_{gj}^3 - \gamma_{gi}^4 - 4\gamma_{gi}^3\gamma_{gj}.
\end{aligned}$$

Now, notice that by replacing whichever of γ_{gi} or γ_{gj} with the larger of the two, we make the negative terms larger. So a sufficient (though again, not necessary) condition for the inequality to be satisfied is:

$$\gamma_{gj}^2\gamma_{gi} \geq 12(\max_k \gamma_{gk})^4,$$

But now notice that this is asking that either

$$(\max_k \gamma_{gk})^2 \min_k \gamma_{gk} \geq 12(\max_k \gamma_{gk})^4,$$

or

$$(\min_k \gamma_{gk})^2 \max_k \gamma_{gk} \geq 12(\max_k \gamma_{gk})^4, \tag{4.11.5}$$

depending on whether $\gamma_{gj} > \gamma_{gj}$ or vice versa. Since we can repeat the logic from equilibrium from Firm j's perspective, we will actually need *both* these conditions to hold for this point to be an equilibrium. But since $\max_{gk} \gamma_{gk} < 1$, it is sufficient that

$$\min_k \gamma_{gk} \geq 12(\max_k \gamma_{gk})^2,$$

which is simply asking that the firms are not too far apart in their marginal costs.

□

Proof of General Tullock case

Our goal is to show the following:

Lemma 4.11.2. *Suppose two firms compete for proportional demand with parameters q and ρ. Suppose further that $\varepsilon^*_{gi} < 1$ for all g and i. If the* nondeviation condition *(as defined below) holds, then both firms playing $(\varepsilon^*_{gi}, \varepsilon^*_{gj})$ is an equilibrium*

$$\varepsilon^*_{gi} = \left(\frac{q\gamma_{gi}}{\rho_g \mu_g}\right)^{\frac{1}{q}} \frac{(\gamma^q_{gi} + \gamma^q_{gj})^{\frac{2}{q}}}{\gamma^q_{gi}\gamma^q_{gj}} = \left(\frac{q}{\rho_g \mu_g}\right)^{\frac{1}{q}} \frac{(\gamma^q_{gi} + \gamma^q_{gj})^{\frac{2}{q}}}{\gamma_{gi}^{1-\frac{1}{q}}\gamma^q_{gj}}.$$

If, furthermore, the investment condition *(as defined below) holds, then this equilibrium is unique.*

Proof. Under the proportional split model of demand, each firm's profit depends not only on its own action, but also that of the other firm. Again, this calls for a game theoretic notion of solution. We look for a pure strategy *Nash Equilibrium*. Recall that in an equilibrium, both firms must be best-responding and have no incentive to deviate.

To find an equilibrium, we first find the best-response of Firm i, *given* the choices of Firm j. Fixing ε_j, the profit of Firm i given the choice of ε is as follows:

$$\pi(\varepsilon_i, \varepsilon_j) = \sum_{g \in \mathcal{G}} \left[\mu_g \frac{\varepsilon^{\rho_g}_{gj}}{\varepsilon^{\rho_g}_{gi} + \varepsilon^{\rho_g}_{gj}}\right] - \sum_{g \in \mathcal{G}} \phi_{gi} + \frac{\gamma_{gi}}{\varepsilon^q_{gi}}.$$

Taking the derivative:

$$\frac{\partial \pi}{\partial \varepsilon_{gi}} = -\mu_g \varepsilon^{\rho_g}_{gj} (\varepsilon^{\rho_g}_{gi} + \varepsilon^{\rho_g}_{gj})^{-2} \left(\rho_g \varepsilon^{\rho_g - 1}_{gi}\right) + \frac{q\gamma_{gi}}{\varepsilon^{q+1}_{gi}}.$$

Setting to zero yields the first-order condition:

$$\frac{q\gamma_{gi}}{\varepsilon^{q+1}_{gi}} = \frac{\rho_g \mu_g \varepsilon^{\rho_g - 1}_{gi} \varepsilon^{\rho_g}_{gj}}{(\varepsilon^{\rho_g}_{gi} + \varepsilon^{\rho_g}_{gj})^2} \implies \frac{\rho_g \mu_g \varepsilon^{\rho_g + q}_{gi} \varepsilon^{\rho_g}_{gj}}{q\gamma_{gi}} = (\varepsilon^{\rho_g}_{gi} + \varepsilon^{\rho_g}_{gj})^2.$$

Applying symmetric logic to Firm j and using the fact that the first order condition for each firm must hold simultaneously in equilibrium, we have that

$$\frac{\rho_g \mu_g \left(\varepsilon^*_{gi}\right)^{\rho_g + q} \left(\varepsilon^*_{gj}\right)^{\rho_g}}{q\gamma_{gi}} = \frac{\rho_g \mu_g \left(\varepsilon^*_{gj}\right)^{\rho_g + q} \left(\varepsilon^*_{gi}\right)^{\rho_g}}{q\gamma_{gj}}.$$

Solving for ε_{gj}^* in terms of ε_{gi}^* yields:

$$\varepsilon_{gj}^* = \varepsilon_{gi}^* \left(\frac{\gamma_{gj}}{\gamma_g i}\right)^{\frac{1}{q}}.$$

Substituting this back in, we have that

$$\begin{aligned}\frac{\rho_g \mu_g \left(\varepsilon_{gi}^*\right)^{\rho_g+q} \left(\varepsilon_{gi}^*\right)^{\rho_g} \left(\frac{\gamma_{gj}}{\gamma_{gi}}\right)^{\frac{\rho_g}{q}}}{q\gamma_{gi}} &= \left(\left(\varepsilon_{gi}^*\right)^{\rho_g} + \left(\varepsilon_{gi}^*\right)^{\rho_g} \left(\frac{\gamma_{gj}}{\gamma_{gi}}\right)^{\frac{\rho_g}{q}}\right)^2 \\ &= \left(\varepsilon_{gi}^*\right)^{2\rho} \left(1 + \left(\frac{\gamma_{gj}}{\gamma_{gi}}\right)^{\frac{\rho_g}{q}}\right)^2.\end{aligned}$$

Solving and rearranging gives that

$$\begin{aligned}\varepsilon_{gi}^* &= \left[\frac{q\gamma_{gi}}{\rho_g \mu_g} \left(\frac{\gamma_{gi}}{\gamma_{gj}}\right)^{\frac{\rho_g}{q}} \left(1 + \left(\frac{\gamma_{gj}}{\gamma_{gi}}\right)^{\frac{\rho_g}{q}}\right)^2\right]^{\frac{1}{q}} \\ &= \left[\frac{q\gamma_{gi}}{\rho_g \mu_g} \left(\frac{\gamma_{gi}}{\gamma_{gj}}\right)^{\frac{\rho_g}{q}} \left(\frac{\gamma_{gi}^{\frac{\rho_g}{q}} + \gamma_{gj}^{\frac{\rho_g}{q}}}{\gamma_{gi}^{\frac{\rho_g}{q}}}\right)^2\right]^{\frac{1}{q}} \\ &= \left[\frac{q\gamma_{gi}}{\rho_g \mu_g} \left(\frac{\gamma_{gi}}{\gamma_{gj}}\right)^{\frac{\rho_g}{q}} \frac{1}{\gamma_{gi}^{\frac{2\rho_g}{q}}} \left(\gamma_{gi}^{\frac{\rho_g}{q}} + \gamma_{gj}^{\frac{\rho_g}{q}}\right)^2\right]^{\frac{1}{q}} \\ &= \left[\frac{q}{\rho_g \mu_g} \frac{\left(\gamma_{gi}^{\frac{\rho_g}{q}} + \gamma_{gj}^{\frac{\rho_g}{q}}\right)^2}{\gamma_{gi}^{\frac{\rho_g}{q}} \gamma_{gj}^{\frac{\rho_g}{q}}}\right]^{\frac{1}{q}}.\end{aligned}$$

Now, notice that the profit can be written as

$$\begin{aligned}
\pi_{gi}(\varepsilon^*_{gi}, \varepsilon^*_{gj}) &= \frac{\left(\frac{q}{\rho_g \mu_g}\right)^{\frac{1}{q}} \frac{(\gamma_{gi}^q + \gamma_{gj}^q)^{\frac{2}{q}}}{\gamma_{gi}^{1-\frac{1}{q}} \gamma_{gj}^q}}{\left(\frac{q}{\rho_g \mu_g}\right)^{\frac{1}{q}} \frac{(\gamma_{gi}^q + \gamma_{gj}^q)^{\frac{2}{q}}}{\gamma_{gj}^{1-\frac{1}{q}} \gamma_{gi}^q} + \left(\frac{q}{\rho_g \mu_g}\right)^{\frac{1}{q}} \frac{(\gamma_{gi}^q + \gamma_{gj}^q)^{\frac{2}{q}}}{\gamma_{gi}^{1-\frac{1}{q}} \gamma_{gj}^q}} - \frac{\gamma_{gi}}{\left(\varepsilon^*_{gi}\right)^q} \\
&= \frac{\left(\frac{1}{\gamma_{gi}^{1-\frac{\rho_g}{q}} \gamma_{gj}^q}\right)^{\rho}}{\left(\frac{1}{\gamma_{gi}^{1-\frac{\rho_g}{q}} \gamma_{gj}^q}\right)^{\rho} + \left(\frac{1}{\gamma_{gj}^{1-\frac{\rho_g}{q}} \gamma_{gi}^q}\right)^{\rho}} - \frac{\gamma_{gi}}{\left(\varepsilon^*_{gi}\right)^q} \\
&= \frac{1}{1 + \left[\frac{\gamma_{gj}^{1-\rho_g/q} \gamma_{gi}^q}{\gamma_{gi}^{1-\rho_g/q} \gamma_{gj}^q}\right]^{\rho_g}} - \frac{\gamma_{gi}}{\left(\varepsilon^*_{gi}\right)^q} \\
&= \frac{1}{1 + \left[\frac{\gamma_{gj}^{1-\rho_g/q-q}}{\gamma_{gi}^{1-\rho_g/q-q}}\right]^{\rho_g}} - \frac{\gamma_{gi}}{\left(\varepsilon^*_{gi}\right)^q}.
\end{aligned}$$

Substituting back in $\left(\varepsilon^*_{gi}\right)^q$, it is:

$$\frac{1}{1 + \left[\frac{\gamma_{gj}^{1-\rho_g/q-q}}{\gamma_{gi}^{1-\rho_g/q-q}}\right]^{\rho_g}} - \frac{\gamma_{gi} \rho_g \mu_g \gamma_{gi}^{\frac{\rho_g}{q}} \gamma_{gj}^{\frac{\rho_g}{q}}}{q \left(\gamma_{gi}^{\frac{\rho_g}{q}} + \gamma_{gj}^{\frac{\rho_g}{q}}\right)^2}.$$

For this interior equilibrium to hold, it must be that $\pi^*_{gi}(\varepsilon^*_{gi}, \varepsilon^*_{gj}) \geq \pi^*_{gi}(\varepsilon', \varepsilon^*_{gj})$ for all other choices ε'. Note that $\pi_{gi,\varepsilon^*_{gj}}(\varepsilon)$ is continuous away from 0. Moreover, for small enough ε, $\pi_{gi,\varepsilon^*_{gj}}(\varepsilon) < 0$, since the market size is bounded by costs can be come arbitrarily negative. Hence, we can consider maximizing this function on the compact set $[\varepsilon_0, 1]$, where ε_0 is the point at which profit becomes negative. Since $\pi_{gi,\varepsilon^*_{gj}}(\varepsilon)$ is continuous on this set, and ε^*_{gi} satisfies the first-order condition, the only possible maxima of this function are ε_0 or 1. At ε_0, the firm is making zero profits, so any choice with positive profits eliminates it. At $\varepsilon = 1$, the firm can also choose to not invest anything in data (and receive the same revenue but no data costs), so the condition that makes $\pi_{gi,\varepsilon^*_{gj}}(\varepsilon^*_{gi}) > \pi_{gi,\varepsilon^*_{gj}}(1)$ will be sufficient to make this an equilibrium.

This condition holds if

$$\frac{1}{1+\left[\frac{\gamma_{gj}^{1-\rho_g/q-q}}{\gamma_{gi}^{1-\rho_g/q-q}}\right]^{\rho_g}} - \frac{\gamma_{gi}\rho_g\mu_g\gamma_{gi}^{\frac{\rho_g}{q}}\gamma_{gj}^{\frac{\rho_g}{q}}}{q\left(\gamma_{gi}^{\frac{\rho_g}{q}}+\gamma_{gj}^{\frac{\rho_g}{q}}\right)^2} \geq \pi_{gi}(1,\varepsilon_{gj^*}). \tag{4.11.6}$$

We call Inequality 4.11.6 the *nondeviation condition*. We can write:

$$\begin{aligned}\pi_{gi}(1,\varepsilon_{gj^*}) &= \frac{\left(\frac{q}{\rho_g\mu_g}\right)^{\frac{1}{q}}\frac{(\gamma_{gi}^q+\gamma_{gj}^q)^{\frac{2}{q}}}{\gamma_{gi}^{1-\frac{1}{q}}\gamma_{gj}^q}}{1+\left(\frac{q}{\rho_g\mu_g}\right)^{\frac{1}{q}}\frac{(\gamma_{gi}^q+\gamma_{gj}^q)^{\frac{2}{q}}}{\gamma_{gi}^{1-\frac{1}{q}}\gamma_{gj}^q}} \\ &= \frac{1}{1+\left(\frac{q}{\rho_g\mu_g}\right)^{-\frac{1}{q}}\frac{(\gamma_{gi}^q+\gamma_{gj}^q)^{\frac{-2}{q}}}{\gamma_{gi}^{\frac{1}{q}-1}\gamma_{gj}^{-q}}},\end{aligned}$$

so Inequality 4.11.6 asks that

$$\begin{aligned}&\frac{1}{1+\left[\frac{\gamma_{gj}^{1-\rho_g/q-q}}{\gamma_{gi}^{1-\rho_g/q-q}}\right]^{\rho_g}} - \frac{\gamma_{gi}\rho_g\mu_g\gamma_{gi}^{\frac{\rho_g}{q}}\gamma_{gj}^{\frac{\rho_g}{q}}}{q\left(\gamma_{gi}^{\frac{\rho_g}{q}}+\gamma_{gj}^{\frac{\rho_g}{q}}\right)^2} \\ &\qquad\geq \frac{1}{1+\left(\frac{q}{\rho_g\mu_g}\right)^{-\frac{1}{q}}\frac{(\gamma_{gi}^q+\gamma_{gj}^q)^{\frac{-2}{q}}}{\gamma_{gi}^{\frac{1}{q}-1}\gamma_{gj}^{-q}}}.\end{aligned}$$

We have shown that if the nondeviation condition holds for each group and each firm, then $(\varepsilon_{gi}^*, \varepsilon_{gj}^*)$ is a Nash Equilibrium in pure strategies under proportionally split demand with competition exponent ρ_g in each group and learning rate q. If a further condition holds, namely that there exists a preferred strategy to non-investment if the opponent invests, then the equilibrium is unique.

Call this the *investment* condition: there exists $\varepsilon \in (0,1)$ such that:

$$\frac{1}{\varepsilon_g^\rho+1} - \frac{\gamma_{gi}}{\varepsilon^q} > \frac{\mu_g}{2} \iff \varepsilon^q - \gamma_{gi}(\varepsilon^{\rho_g}+1) \geq \frac{\mu_g\left((\varepsilon^{\rho_g}+1)\varepsilon^q\right)}{2}. \tag{4.11.7}$$

Equivalently, we need to ensure that there is an $\varepsilon \in (0,1)$ such that:

$$\iff \varepsilon^q - \gamma_{gi}(\varepsilon^{\rho_g}+1) - \frac{\mu_g\left((\varepsilon^{\rho_g}+1)\varepsilon^q\right)}{2} \geq 0 \tag{4.11.8}$$

has a solution in $(0, 1)$. This will not always be the case, of course; when it is not, then there is an equilibrium in which both firms prefer not to invest in collecting data from one group at all, which certainly exacerbates inequality. □

4.12 Omitted Proofs from Section 4.6.1

Omitted Algebra for Optimal Profit. Recall that we would like to show that the optimal profit achievable by the monopolist facing parameters $\mu_g, \alpha_g, \beta_g, \gamma_g$ for g in $\mathcal{G} = \{A, B\}$ is:

$$\pi^*(\varepsilon^*) = \sum_{g \in \{A,B\}} \mu_g \alpha_g - (\mu_g \beta_g)^{\frac{q}{q+1}} \gamma_g^{\frac{1}{q+1}} Q.$$

For clarity, write η_g for $\mu_g \beta_g$. Then we can write the optimal profit for a group as a function of the parameters using the result that the profit optimizing choice of error is $\varepsilon_g^* = (q\gamma_g/\eta_g)^{1/(q+1)}$.

$$\begin{aligned}
\pi_g^*(\mu_g, \gamma_g, \alpha_g, \beta_g) &= \alpha_g \mu_g - \eta_g \varepsilon_g^* - \gamma / \varepsilon_g^{*q} \\
&= \alpha_g \mu_g - \eta_g \eta_g^{-\frac{1}{q+1}} q^{\frac{1}{q+1}} \gamma_g^{\frac{1}{q+1}} - \gamma_g \left(\frac{\eta_g}{q\gamma_g} \right)^{\frac{q}{q+1}} \\
&= \alpha_g \mu_g - \eta_g^{\frac{q}{q+1}} \gamma_g^{\frac{1}{q+1}} q^{\frac{1}{q+1}} - \gamma_g^{\frac{1}{q+1}} \frac{1}{q^{q/(q+1)}} \eta_g^{\frac{q}{q+1}} \\
&= \alpha_g \mu_g - \eta_g^{\frac{q}{q+1}} \gamma_g^{\frac{1}{q+1}} \left[q^{\frac{1}{q+1}} + \frac{1}{q^{q/(q+1)}} \right].
\end{aligned}$$

Then writing $Q = q^{\frac{1}{q+1}} + \frac{1}{q^{q/(q+1)}}$, substituting back $\mu_g \beta_g$ for η_g, and summing over groups yields the claim. □

Proof of Lemma 4.6.2. Fix a solution $(\varepsilon_A, \varepsilon_B)$ to the constrained profit optimization problem. We will show that unless $\varepsilon_B = (1 + \chi)\varepsilon_A$, $(\varepsilon_A, \varepsilon_B)$ is not a constrained profit maximizer.

Since by assumption $\varepsilon_B^M > \varepsilon_A^M(1+\chi)$ but $\varepsilon_A/(1+\chi) \leq \varepsilon_B \leq (1+\chi)\varepsilon_A$, we can't have both $\varepsilon_B^M = \varepsilon_B$ and $\varepsilon_A^M = \varepsilon_A$. Without loss of generality, assume that $\varepsilon_B^M \neq \varepsilon_B$.

There are three cases. In the first case, $\varepsilon_B^M > (1+\chi)\varepsilon_A$. We can increase the profit achieved by $(\varepsilon_A, \varepsilon_B)$ by increasing ε_B, as in this case, $\varepsilon_B < \varepsilon_B^M$. To see this, let $\varepsilon_\alpha = \left(\varepsilon_A, \alpha\varepsilon_B^M + (1-\alpha)\varepsilon_B\right)$ for $\alpha \in [0,1]$. By Jensen's inequality, there is an α such that

$$\pi(\varepsilon_\alpha) \geq (1-\alpha)\pi\left((\varepsilon_A, \varepsilon_B)\right) + \alpha\pi\left(\left(\varepsilon_A, \varepsilon_B^M\right)\right) > \pi((\varepsilon_A, \varepsilon_B)).$$

The first inequality holds for any $\alpha \in [0,1]$, so we set α so that $\varepsilon_\alpha = (\varepsilon_A, (1+\chi)\varepsilon_A)$, in which case this is still a feasible solution, and by the separability of the profit function, the second inequality holds.

In the second case, $\varepsilon_B^M < \varepsilon_A/(1+\chi)$. Then by the same logic using Jensen's inequality, we can increase the profit by decreasing ε_B to $\varepsilon_A/(1+\chi)$, i.e. $\pi((\varepsilon_A, \varepsilon_B)) < \pi((\varepsilon_A, \varepsilon_A/(1+\chi)))$. But we can increase the profit even more in this case because $\varepsilon_A^M < \varepsilon_B^M/(1+\chi) < \varepsilon_A/(1+\chi)^2$, so now we can decrease ε_A to see that profit is maximized at $(\varepsilon_A/(1+\chi)^2, \varepsilon_A/(1+\chi))$.

Otherwise, $\varepsilon_A/(1+\chi) \leq \varepsilon_B^M \leq (1+\chi)\varepsilon_A$. This is very similar to the previous case: Jensen's inequality along with the separability of π shows that $\pi((\varepsilon_A, \varepsilon_B)) < \pi\left(\left(\varepsilon_A, \varepsilon_B^M\right)\right) \leq \left(\varepsilon_B^M/(1+\chi), \varepsilon_B^M\right)$.

□

Proof of Corollary 4. First, we show that $\varepsilon_A^R \geq \varepsilon_A^M$:

$$\begin{aligned}
\left(\frac{\varepsilon_A^R}{\varepsilon_A^M}\right)^{q+1} &= \frac{q\left(\gamma_A + \gamma_B/(1+\chi)^q\right) / \left(\mu_A\beta_A + \mu_B\beta_B(1+\chi)\right)}{q\gamma_A/(\mu_A\beta_A)} \\
&= \frac{\mu_A\beta_A}{\mu_A\beta_A + \mu_B\beta_B(1+\chi)} \frac{\gamma_A + \gamma_B/(1+\chi)^q}{\gamma_A} \\
&= \frac{\mu_A\beta_A\gamma_A + \mu_A\beta_A\gamma_B/(1+\chi)^q}{\mu_A\beta_A\gamma_A + \mu_B\beta_B\gamma_A(1+\chi)}.
\end{aligned}$$

Notice that

$$\frac{\varepsilon_A^R}{\varepsilon_A^M} \geq 1 \iff \left(\frac{\varepsilon_A^R}{\varepsilon_A^M}\right)^{q+1} \geq 1.$$

Now using the elementary fact that for positive x, y, z, $(x+y)/(x+z) \geq 1 \iff y \geq z$, we can see that

$$\left(\frac{\varepsilon_A^R}{\varepsilon_A^M}\right)^{q+1} \geq 1 \iff \mu_A\beta_A\gamma_B/(1+\chi)^q \geq \mu_B\beta_B\gamma_A(1+\chi)$$
$$\iff \frac{\gamma_B}{\mu_B\beta_B} \geq \frac{\gamma_A}{\mu_A\beta_A}(1+\chi)^{q+1}.$$

But recalling that the monopolist's optimal solution is $\varepsilon_g^M = (\frac{q\gamma_g}{\mu_g\beta_g})^{\frac{1}{q+1}}$, we can rewrite the previous inequality as

$$\varepsilon_B^{Mq+1} \geq \varepsilon_A^{Mq+1}(1+\chi)^{q+1} \iff \varepsilon_B^M \geq \varepsilon_A^M(1+\chi),$$

which is exactly Assumption 1.

Now, we show that $\varepsilon_B^R \leq \varepsilon_B^M$:

$$\varepsilon_B^R = (1+\chi)\varepsilon_A^R = \left(q\frac{((1+\chi)^{q+1}\gamma_A + \gamma_B(1+\chi))}{\mu_A\beta_A + \mu_B\beta_B(1+\chi)}\right)^{\frac{1}{q+1}}.$$

Then

$$\left(\frac{\varepsilon_B^R}{\varepsilon_B^M}\right)^{q+1} = q\frac{\gamma_A(1+\chi)^{q+1} + \gamma_B(1+\chi)}{\mu_A\beta_A + \mu_B\beta_B(1+\chi)}\frac{\mu_B\beta_B}{q\gamma_B}$$
$$= \frac{\mu_B\beta_B\gamma_A(1+\chi)^{q+1} + \mu_B\beta_B(1+\chi)\gamma_B}{\gamma_A\mu_A\beta_A + \mu_B\beta_B\gamma_B(1+\chi)}.$$

So again using the elementary fact that $\frac{y+x}{z+x} \iff y < z$, we have:

$$\left(\frac{\varepsilon_B^R}{\varepsilon_B^M}\right)^{q+1} \leq 1 \iff \mu_B\beta_B\gamma_B(1+\chi) \leq \mu_A\beta_A\gamma_A$$
$$\iff \frac{\gamma_B}{\mu_B\gamma_B} \geq \frac{\gamma_A}{\mu_A\beta_A}(1+\chi)^{q+1}$$
$$\iff \varepsilon_B^{Mq+1} \geq \varepsilon_A^{Mq+1}(1+\chi)^{q+1}$$
$$\iff \varepsilon_B^M \geq \varepsilon_A^M(1+\chi).$$

□

Proof of Corollary 5. Returning to the second equation in the proof of Corollary 4 :

$$\begin{aligned}(\text{PoF}_{1+\chi})^{q+1} &= \frac{\mu_A\beta_A}{\mu_A\beta_A+\mu_B\beta_B(1+\chi)}\frac{\gamma_A+\gamma_B/(1+\chi)^q}{\gamma_A}\\ &\leq \frac{\gamma_A+\gamma_B/(1+\chi)^q}{\gamma_A}.\end{aligned}$$

Taking the $(q+1)$th-root yields the claim. □

Proposition 13. We can write

$$\text{MonPoF}_{1+\chi} = \frac{\mu_A\alpha_A + r\mu_A\alpha_B - Q\mu_A^{\frac{q}{q+1}}\left[\beta_A^{\frac{q}{q+1}}\gamma_A^{\frac{1}{q+1}} + r^{\frac{q}{q+1}}\beta_B^{\frac{q}{q+1}}\gamma_B^{\frac{1}{q+1}}\right]}{\mu_A\alpha_A + r\mu_A\alpha_B - Q\mu_A^{\frac{q}{q+1}}\left[(\beta_A+r\beta_B(1+\chi))^{\frac{q}{q+1}}(\gamma_A+\gamma_B\frac{1}{(1+\chi)^q})^{\frac{1}{q+1}}\right]}$$

Factoring out μ_A and using the fact that if $\mu_B \to \infty$, $\mu_A \to \infty$, we have:

$$\begin{aligned}\lim_{\mu_B\to\infty}\text{MonPoF}_{1+\chi} &= \lim_{\mu_A\to\infty}\text{MonPoF}_{1+\chi}\\ &= \lim_{\mu_A\to\infty}\frac{\alpha_A + r\alpha_B - Q\mu_A^{-\frac{1}{q+1}}\left[\beta_A^{\frac{q}{q+1}}\gamma_A^{\frac{1}{q+1}} + r^{\frac{q}{q+1}}\beta_B^{\frac{q}{q+1}}\gamma_B^{\frac{1}{q+1}}\right]}{\alpha_A + r\alpha_B - Q\mu_A^{-\frac{1}{q+1}}\left[(\beta_A+r\beta_B(1+\chi))^{\frac{q}{q+1}}(\gamma_A+\gamma_B\frac{1}{(1+\chi)^q})^{\frac{1}{q+1}}\right]}\end{aligned}$$

But then

$$\lim_{\mu_B\to\infty}\text{MonPoF}_{1+\chi} = 1,$$

since $\mu_A^{-1/(q+1)} \to 0$ as $\mu_A \to \infty$ and its multipliers are constants.

For the second claim, we can simply substitute in $\mu_B = 0$ and factor out $\mu_A\alpha_A$ to get

$$\begin{aligned}\lim_{\mu_B\to 0}\text{MonPoF} &= \frac{\mu_A\alpha_A[1-Q\left(\frac{\gamma_A}{\mu_A\alpha_A}\right)^{\frac{1}{q+1}}]}{\mu_A\alpha_A[1-Q\left(\frac{\gamma_A+\gamma_B/(1+\chi)^{\frac{1}{q+1}}}{\mu_A\alpha_A}\right)^{\frac{1}{q+1}}]}\\ &= \frac{[1-Q\left(\frac{\gamma_A}{\mu_A\alpha_A}\right)^{\frac{1}{q+1}}]}{1-Q\left[\frac{\gamma_A+\gamma_B/(1+\chi)^{\frac{1}{q+1}}}{\mu_A\alpha_A}\right]^{\frac{1}{q+1}}}\\ &= \frac{1-Q/q\varepsilon_A^M}{1-Q/q\varepsilon_A^M\left(1+\frac{\gamma_B}{\gamma_A}\frac{1}{(1+\chi)^{\frac{1}{q+1}}}\right)^{\frac{1}{q+1}}}.\end{aligned}$$

□

4.13 Omitted Proofs from Section 4.6.2

Proof of Lemma 4.6.5. First, note that the absolute error constraints are separable, so that the firm's profit maximization problem is simply $\max_{\varepsilon_g} \pi_g(\varepsilon_g)$ subject to $\varepsilon_g \leq \chi$, for each of $g \in \{A, B\}$. So fix a group $g \in \{A, B\}$. If $\varepsilon_g^R \neq \varepsilon_g^M$, it must be that $\varepsilon_g^M > \chi$, as otherwise the constraint would have already been met by ε_g^M. Now we show that for any feasible error rate $\varepsilon_g \leq \chi$, $\pi_g(\varepsilon_g) \leq \pi_g(\chi)$ with equality holding only at $\varepsilon_g = \chi$. So suppose that $\varepsilon_g < \chi$.

Let $\varepsilon_{g,\alpha} = (1-\alpha)\varepsilon_g + \alpha\varepsilon_g^M$ for $\alpha \in [0, 1]$. Since π_g is concave, Jensen's inequality implies that

$$\begin{aligned}\pi_g(\varepsilon_{g,\alpha}) &= \pi_g\left((1-\alpha)\varepsilon_g + \alpha\varepsilon_g^M\right) \\ &\geq (1-\alpha)\pi_g(\varepsilon_g) + \alpha\pi_g\left(\varepsilon_g^M\right) \\ &> \pi_g(\varepsilon_g),\end{aligned}$$

with the second inequality holding when $\alpha > 0$ because $\pi_g\left(\varepsilon_g^M\right) > \pi_g(\varepsilon_g)$. Then choosing $\alpha = \frac{\chi - \varepsilon_g}{\varepsilon_g^M - \varepsilon_g}$ suffices as then $\varepsilon_{g,\alpha} = \chi$. (Notice that $\frac{\chi-\varepsilon_g}{\varepsilon_g^M-\varepsilon_g} \in (0, 1]$ since $\varepsilon_g^M > \chi$ and $\chi > \varepsilon_g$.)

□

Proof of Proposition 14. First, by Lemma 4.6.5, when $\chi < \varepsilon_A^M$, then the optimal choice for the monopolist is $(\varepsilon_A^R, \varepsilon_B^R) = (\chi, \chi)$. In this case, the profit is

$$\pi(\chi, \chi) = \mu_A(\alpha_A - \beta_A\chi) + \mu_B(\alpha_B - \beta_B\chi) - \phi_A - \phi_B - \frac{\gamma_A}{\chi^q} - \frac{\gamma_B}{\chi^q}.$$

Since π is concave, the global optimum is at $(\varepsilon_A^M, \varepsilon_B^M)$, and as the error rate goes to zero, profit goes to negative infinity, there must be a minimum error rate $\chi_0 > 0$ where the profit is zero. Since this error rate must be smaller than ε_A^M, the above formula for profit holds and this error rate is the solution to

$$\mu_A(\alpha_A - \beta_A\chi) + \mu_B(\alpha_B - \beta_B\chi) - \phi_A - \phi_B - \frac{\gamma_A}{\chi^q} - \frac{\gamma_B}{\chi^q} = 0.$$

Multiplying by χ^q and re-arranging gives the claim. □

Proof of Corollary 15. For $\mu_B \to \infty$, note that we can write the price of fairness as:

$$\begin{cases} 1 & \chi \geq \varepsilon_B^M \\ \frac{\pi_A(\varepsilon_A^M)+\pi_B(\varepsilon_B^M)}{\pi_A(\varepsilon_A^M)+\pi_B(\chi)} & \varepsilon_A^M \leq \chi < \varepsilon_B^M \\ \frac{\pi_A(\varepsilon_A^M)+\pi_B(\varepsilon_B^M)}{\pi_A(\chi)+\pi_B(\chi)} & \chi_0 < \chi < \varepsilon_A^M \\ \infty & \chi < \chi_0 \end{cases}.$$

Since $\varepsilon_A^M, \varepsilon_B^M \to 0$ as $\mu_B, \mu_A \to \infty$, as the population grows, eventually ε_B^M and ε_A^M will be less than χ, so that $\varepsilon_B^R = \varepsilon_B^M$ and $\varepsilon_A^R = \varepsilon_A^M$. Thus $\lim_{\mu_B \to \infty} \text{MonPoF}_\chi = 1$.

For $\mu_B \to 0$, given our assumption on χ, we will be be in either Case 1, Case 2, or Case 3. Note that as $\mu_B \to 0$, ε_B^M will eventually be larger than χ, so the limit will be obtained at either Case 2 or Case 3. In Case 2, we can substitute in 0 for μ_B; combining this with the fact that for small enough μ_B, the optimal choice for the unconstrained monopolist eventually becomes to set $\varepsilon_B^M = 1$, we can write the price of fairness as:

$$\begin{aligned} \lim_{\mu_B \to \infty} \text{MonPoF}_\chi &= \frac{\pi_A(\varepsilon_A^M) - \gamma_B - \phi_B}{\pi_A(\varepsilon_A)^M - \gamma_B/\chi^q - \phi_B} \geq 1 \\ &= \frac{\mu_A\alpha_A - Q\gamma_A^{\frac{1}{q+1}}(\mu_A\beta_A)^{\frac{q}{q+1}} - \phi_A - \gamma_B - \phi_B}{\mu_A\alpha_A - Q\gamma_A^{\frac{1}{q+1}}(\mu_A\beta_A)^{\frac{q}{q+1}} - \phi_A - \gamma_B/\chi^q - \phi_B} \\ &= \frac{1 - \frac{\gamma_B - \phi_B}{\frac{1}{q+1}(\mu_A\beta_A)^{\frac{q}{q+1}} - \phi_A}}{1 - \frac{\gamma_B/\chi^q - \phi_B}{\frac{1}{q+1}(\mu_A\beta_A)^{\frac{q}{q+1}} - \phi_A}} \geq 1. \end{aligned}$$

since $\chi \leq 1 \implies \gamma_B/\chi^q \geq \gamma_B$.

Alternatively, if Case 3 obtains, then we can write:

$$\lim_{\mu_B \to \infty} \text{MonPoF}_\chi = \frac{\pi_A(\chi) - \gamma_B - \phi_B}{\pi_A(\chi) - \gamma_B/\chi^q - \phi_B} \geq 1.$$

□

Chapter 5

FAIRNESS IN THE MORTGAGE MARKET

5.1 Introduction[1]

In the final chapter of this thesis, we apply notions of fairness from the Fair Machine Learning (Fair ML) literature to one of the most consequential markets for individuals: the mortgage market of the United States. Our broad motivation can be easily seen in a picture: Figure 5.1 displays the denial rates and ultimate default rates of first lien mortgages for purchase of single family homes broken down by racial group over time. It is clear that, despite clear macroeconomic influences and fluctuations over time, there is a persistent gap in denial rates between majorities and minorities, and this is particularly pronounced for Black Americans. Against a backdrop of historical and present racial inequality, these figures on mortgages – a channel that often provides for intergenerational wealth transition and wealth

[1]This Chapter is based on work conducted while the author was a part-time employee of the Federal Reserve Bank of Philadelphia with Simon Freyaldenhoven and Minchul Shin. The views expressed in this work are solely those of the author and do not represent those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System. All models are strictly academic and not for commercial use.

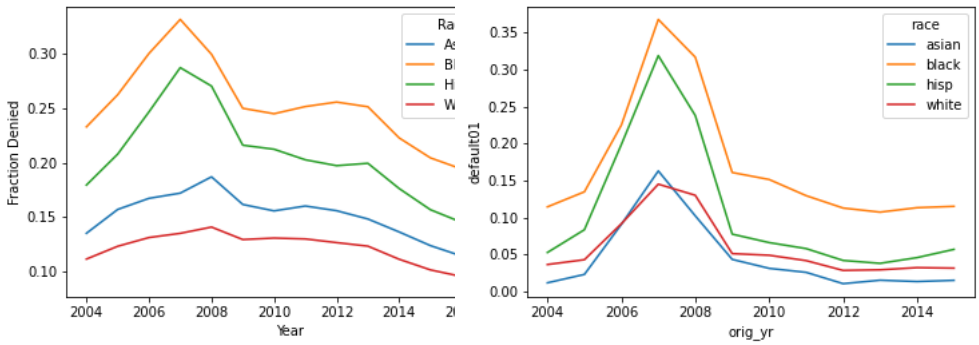
creation – represent a stark picture of a progress yet to be made.

Our goal in this work is twofold. First: this market historically harbored a great deal of discrimination that resulted in significant inequity; as such, it has been the subject of both activist movements and legislative intervention, and an understanding of to what extent problems are continuing to be perpetuated (as opposed to inherited) is an important input to policy. Thus, as a policy matter, we would like to provide some measure of just *how* fair this market is as a way of measuring the progress made. Furthermore, we would like to identify and quantify various *counterfactual* policy choices and what their implications would be on the fairness measures we evaluate; in doing so, we elucidate the policy trade-offs involved in choosing a policy regime.

Second, we view this market as a particularly appealing laboratory in which to apply these fairness measures in a new and much more complicated setting. Most of Fair ML has developed in the setting of a single learner's attempt to be fair while training a model on their own data. We ask whether this literature can be adapted for a broader perspective. This is, of course, a reductive portrayal, and the field has certainly been expanding into new settings, notions, and directions; however, the complexity of the mortgage market, and the range of technical challenges in analyzing this market in practice, appear to be unique to date in the fairness literature.

Our goal is thus to *empirically* measure fairness in this market using quantitative fairness notions. The mortgage market, like the abstract data-driven market we studied in Chapter 4, is a market that depends on predictions gleaned from data, but it is of course much more complicated: even the *definition*, let alone *measurement*, of fairness in this setting is unclear, and can be fraught with pitfalls. In this chapter, we present our approach and initial results to address this question using a unique, large dataset of mortgage outcomes and applications.

We follow and expand on the content of ongoing work presented at [INFORMS



(a) Denial rates by application year and race.

(b) Defaults by origination year and race.

Figure 5.1: Aggregates by Race. Source: HMDA.

presentation]. Our organization is as follows. The rest of this section provides some background on mortgages and overviews related work with a focus on the economics literature, while we overview fairness more broadly in Section 5.2. Section 5.3 describes our approach, including our modeling of fairness regime and predictions and justification for our focus on marginal applicants. We give our results in Section 5.4.

5.1.1 Mortgages

A *mortgage* is commonly defined as a debt instrument secured by real estate, most often used to purchase property with the collateral being the property itself. Today, the thirty-year fixed loan is perhaps the most commonly used [92], though pre-great depression, the most common structure was interest-only payments for five years that "ballooned" into the full balance of the loan at the end. The mortgage market is consequential for several reasons, the most salient of which is that debt-financed real estate investment has been an important channel for asset accumulation and intergenerational wealth transfer among the middle and upper classes. It has thus been seen as vehicle for social mobility and thus encouraged by government pol-

icy. And symbolically, owning one's home has been seen as the American dream. Historically, the mortgage market has suffered from significant discrimination. Restrictive covenants, eventually declared legally unenforceable, restricted the sale of some homes to minorities. Redlining, the practice of using geographic characteristics to deny loans in largely minority areas, ostensibly to avoid property value risk, greatly limited minority home ownership. Other practices, including exclusionary zoning, promoted racial segregation. See [110] for a fuller accounting. As a consequence, fair housing was a significant component of activist demands in the Civil Rights movement, culminating in the Fair Housing Act (as part of the Civil Rights Act of 1968) and later legislative actions which similarly intended to mitigate housing discrimination (including the Home Mortgage Disclosure Act and Equal Opportunity Credit Act).

Since the great depression, the government has taken an active role in the mortgage market, including via insuring mortgages or guaranteeing mortgages via government-sponsored enterprises (GSEs) like the Federal National Mortgage Association (FNMA or Fannie Mae) [cite]. The purposes of these interventions and the relationship to racial equality has depended on the historical period; [110] argues that federal loan insurance in the first half of the 20th century had discriminatory intent built-in by design via redlining[2], and [118] argues that even after discrimination in the housing market was formally ended various government-encouraged practices continued to promote discrimination; more recently, on the other hand, the federal government has applied policies designed to promote home ownership among minorities, including pressure on GSEs to relax credit and down-payment standards; [3] argues that this may be one of the factors[3] that ultimately led to the financial crisis and Great Recession of 2008.

[2]Others have argued that redlining reflected existing discrimination rather than intentionally promoting it [57]; regardless, the effect seems to have been to exacerbate segregation.

[3]This argument is controversial, however. [5] argues that the Community Reinvestment Act, one particular policy, led to riskier lending, while [18] rebuts this claim.

There are many types of mortgages with various details, but an important high-level division is between *conventional* and *unconventional* mortgages. Conventional loans are loans that are not directly guaranteed by government agencies, as opposed to *unconventional* loans which are. The difference may be less clear in practice, as the majority of conventional loans are *conforming* loans (that is, conform to GSEs' requirements for sale eligibility) and subsequently sold off to GSEs. However, between non-conforming loans which cannot be sold off, and the risk of put-backs[4] ensure that the conventional loans tend to have more stringent requirements.

A crucial outcome that we will look at is *default*. Mortgages have specified payment schedules with interest rates that may be fixed or vary with marketwide interest rates; if a payment is not made by the specified time or amount that it is due, a loan becomes delinquent. Technically, the loan is *in default*, but because recovering the full value from a loan in default is difficult, banks often would prefer a delinquent loan to return to current status rather than engage in foreclosure and recovery proceedings. Thus practically speaking, default is better defined as being in delinquency for a period of time long enough that banks tend to begin recovery proceedings; often, in the literature, this is taken to be 90 days, and that is the standard we will use.

5.1.2 Related Work

In this subsection we briefly discuss empirical and theoretical work in social science that addresses discrimination and the mortgage market; we explore work relating to fairness more broadly in Section 5.2.

In economics, perhaps the first and certainly the most famous formal treatment of discrimination was given by Becker in [13], grounded in the labor market, which

[4]Put-backs are when the the GSE discovers that the loan in question does not, in fact, conform, due to an error or oversight; the GSE is then able to "put back" the loan and the risk remains on the originator's books.

modeled discrimination as an "animus" on the part of some employers. In this setting, employers or their employees suffer a disutility from employing or working with minorities, and so will only hire them at an effectively lower wage given their productivity. This form of animus would, in theory, be competed away in a competitive enough market. An alternative notion, often taken as a starting point by much of the discrimination research (in Economics), is called *statistical* discrimination, pioneered by Arrow [7] and Phelps [104]. In statistical discrimination, decision-makers are not motivated by any animus, but use group membership as a signal in the face of incomplete information to maximize profits. This notion of discrimination, while perhaps less morally odious than animus-based discrimination, is little fairer to those discriminated (and in general, equally illegal).

There are several recent works that are also interested in specifically the question of discrimination in mortgages. [12], for instance, evaluates discrimination through the channel of pricing (here, interest rates) by focusing on mortgages with no credit risk (since they are immediately sold off to GSEs) and examining pricing differential by race and ethnicity. They find that Latino and Black borrowers pay in aggregate an extra $765MM in interest per year relative to Whites, but that discrimination is 40% lower in FinTech lenders (which rely heavily on algorithms). Another related paper is [60], which focuses on how improved predictive technology will impact the *distribution* of errors, and how they may disproportionately affect minority groups. The authors build an equilibrium model and then test the prediction in practice on mortgage data, and find that machine learning improvements seem to increase disparities across groups. Both [12] and [60] take a similar system-wide perspective as we do in the chapter, and rely on similar datasets. On the other hand, [41] is perhaps more similar to our work in the spirit of detecting discrimination in loan decision-making; they apply what amounts to the outcome test (described in [13] and relative to predictive parity in Section 5.2.1 to administrative data from a single high-cost lender, and exploit quasi experimental variation in loan officer assignment

to find bias in lending to immigrants and older applicants.

5.2 Measuring Fairness: Definitions and Impossibilities

The immediate question that occurs is: what does it mean to be fair? This is, of course, a philosophical question, not a mathematical one, and indeed, a question that dates back to the very beginnings of philosophy. In this work, we will ultimately not take a position on what the "true" meaning of fairness is (and certainly not hope to provide a complete survey on the various ideas that have been adopted) but we must at least establish a framework for what we are to measure.

There have been many fairness settings studied – far too many to list here[5] – and notions of fairness explored, and algorithms created; at this point, the literature is quite rich and relatively mature. The study of fairness in machine learning has been spurred by the empirical discovery of apparently unfair results in state of the art systems of various levels of stakes (e.g., [126], [25], [21]), and has produced new methods in a variety of settings including supervised learning [4], online and reinforcement learning [76],[75], [51], and representation learning [127]. One important set of fairness notions that are widely used is the family of group fairness definitions like statistical parity [44], equality of opportunity [67], equalized odds [67], calibration [105], and so on.

As pointed out by [71], many of these definitions can be interpreted from a perspective of (various notions of) *equality of opportunity*. In particular, many of the common fairness notions can be viewed as an instantiation of the Rawlsian "veil of ignorance" perspective: we view the model as apportioning a good or bad, which may be the label itself (e.g the loan) or the quality of the label (e.g falsely denying a loan), and so on; fundamentally, these should not depend on irrelevant

[5] See [10] for a comprehensive introduction to many aspects of fairness in machine learning.

characteristics to the task at hand, like race or gender. Then, we ask how we would like the model to act if we were behind a veil of ignorance – that is, without knowing what member of society we will be or what group we will belong to. Behind a veil of ignorance, models that are equally likely to bestow the good regardless of group status seem preferable to those that do not. Possibly the most influential philosopher in very recent times [22], Rawls' version of "justice as fairness" [106], and his thoughts on justice and fairness have been very influential on the legal system, which is where most fairness concerns will ultimately be decided.

Much of the legal foundations of anti-discrimination come from statutory law like the Fair Housing Act, the American Disabilities Act, or other titles of the Civil Rights Act; the Equal Protection Clause of the Constitution is, of course, also foundational. The two common legal theories used in anti-discrimination jurisprudence are *disparate treatment* and *disparate impact*. Broadly, disparate treatment can be thought of as the intent to discriminate, including explicitly using protected attributes in decision-making; disparate impact occurs when decision-makers do not explicitly use protected attributes, or intend to discriminate, but nonetheless their decision-making causes outcomes that differ greatly by protected attribute.

The following from [11] describes clearly these two theories of discrimination:

> *Disparate treatment* comprises two different strains of discrimination: (1) formal disparate treatment of similarly situated people and (2) intent to discriminate. *Disparate* impact refers to policies or practices that are facially neutral but have a disproportionately adverse impact on protected classes. [emphasis added]

Note that taking into account protected attributes is not always illegal, nor are processes that result in differential hiring rates. For example, taking into account protected attributes in order to rectify historical inequities is permissible; indeed, this is legally-recognized affirmative action. Some processes that result in adverse impact can be justified by business necessity. The actual determination of the

illegality of an action is governed in some cases by explicit statutory law, but often requires litigation and interpretation of particular circumstances by courts.

In Section 5.3, we will conceptualize stylized policy regimes of *No Disparate Treatment* and *No Disparate Impact* as constraints on the action space of a bank that wishes to maximize its profits. For No Disparate treatment, we imagine that banks must choose thresholds that are blind to race; for No Disparate Impact, we imagine that banks must lend out at equal rates to different groups (and so, maximizing profit, will prefer to lend to the top part of the distribution). The models we use are very simple, and of course these regimes are much more complicated in real life. Still, we will see that the models do seem to serve as a useful conceptual device, and make testable predictions about the various fairness quantities defined in Section 5.2.1. Hence, these regimes and models will guide our analysis.

5.2.1 Specific Definitions

In discussing these definitions, we will ground them in our particular context - loans and racial discrimination - and so use the following notation. The learner's goal is to predict whether an applicant will profitably repay a loan (y) using some features X. We will assume the decision maker also has access to a protected class membership, g, which they can use to evaluate their model for fairness purposes and which they may or may not be able to use in making decisions (though often, in practice, this would be illegal). Many treatments write $\hat{y}$ for our prediction of y, but we will write ℓ to emphasize the fact that a loan is being allocated to those whom we predict positively. We will write d for the complement of y - that is, $d = 1$ if the borrower defaults and 0 otherwise. For each definition, we will treat all outcomes as binary.

The first definition is possibly the most natural:

Definition 5.2.1 (Demographic Parity)**.** A binary classifier satisfies *demographic*

parity (DP) if, for any G, G':

$$\Pr[\ell = 1 | g = G] = \Pr[\ell = 1 | g = G']$$

Note that we can also write DP in terms of expectations of ℓ: $\mathrm{E}[\ell | g = G] = \mathrm{E}[\ell | g = G']$. Notice that if underlying base rates are not equal, a perfectly accurate classifier would not satisfy demographic parity, and so satisfying DP even with perfect information would require making mistakes in terms of prediction. In our context, that would require the bank to make some loans it knows to be a loss.

A more satisfying definition (and arguably more in line with Rawlsian ideas) is the following:

Definition 5.2.2 (Equality of Opportunity)**.** A binary classifier satisfies *equality of opportunity* (EO) if, for any G, G':

$$\Pr[\ell = 1 | d = 0, g = G] = \Pr[\ell = 1 | d = 0, g = G']$$

Unfortunately, EO cannot be easily measured in a setting such as ours because of the *selective labels* problem. That is, whether an applicant was wrongly denied a loan cannot be directly identified,[6] because there was no loan made on which to default. To avoid this issue, measuring EO for a given decision policy requires a period of making loans to all applicants (against the policy's prescription) in order to evaluate what the results of the policy *would have been* had it been followed. Of course, one may approximate this by experiment, e.g. following the policy for the most part but randomizing making a loan when denial is recommended.

An alternative is Conditional Demographic Parity (CDP):

Definition 5.2.3 (Conditional Demographic Parity)**.**

CDP effectively measures demographic parity *given* feature attributes. Why should we consider it a conciliatory replacement for EO? An argument that can

[6] Unless they make multiple applications to different banks; we explore this in future work.

be made formal is the following: suppose that risk is truly a function of some set of features of the applicants. Then if one had that function and those features, one could directly compute default probability, and then recover a measurement of EO by evaluating decisions against calculated risk level. Of course, one does not have that underlying function. But with a complicated enough model class and enough data, then as more features[7] are added into a learning process, the limit will approximate the true risk model. In this work, however, we will not consider EO or CDP in detail.

A third commonly used fairness notion is *positive predictive value* (PPV):

Definition 5.2.4 (Equal PPV)**.** We say that a binary classifier satisfies *equalized positive predictive value* if for any G, G':

$$\Pr[d = 1 | \ell = 1, g = G] = \Pr[d = 1 | \ell = 1, g = G']$$

Requiring a decision rule to satisfy equal PPV is tantamount to the *outcome test* proposed in the statistical discrimination literature. The reasoning behind the outcome test is as follows: suppose that some group is defaulting at a much lower rate than others. Then it would appear that this group must be more creditworthy to be approved, suggesting that the decision-maker is holding them to a higher standard. There are problems with this as a be-all, end-all test: for instance, the problem of *inframarginality* [115] can result in PPV appearing unequal in the absence of discrimination if risk distributions are different above the threshold. (We can, and do, mitigate this issue somewhat by focusing on candidates around the margin, but this is not the only issue with PPV.)

These fairness measures, among others, have clear interpretations in terms of the quantities they represent, and adopting any of them as an exclusive fairness

[7] Here, we again are subject to the specter of omitted variable bias. But importantly, one could learn this model with a set of experiments or loans to all applicants and thus be able to get at EO rather than not being able to measure EO at all.

measure emphasizes some quantities over others. These also implicitly emphasize different sorts of harms, goods, and actors, and choosing between them amounts to making moral or philosophical choices. Unfortunately, any decision rule, except in very special cases, *cannot* simultaneously satisfy all non-discrimination criteria, as researchers discovered [34], [86] in responding to a significant and illustrative controversy about algorithmic decision tools[8]. Hence, there should not be a hope that we, as society, can avoid taking a stand on the these moral questions. Yet to do so well, we should understand the trade-offs we face; that is what this work hopes to elucidate.

5.3 Framework

5.3.1 Profit and policy

We consider three potential policy regimes. The first is an *unconstrained* regime in which banks may do whatever they like, including discriminating based on race, in an attempt to maximize profit (without any inherent racial animus). This is, of course, against anti-discrimination law, but serves as a useful benchmark. The other regimes are given by two extremes based on disparate treatment and disparate impact law, interpreted literally. That is, in the second regime (*No Disparate Treatment*), we suppose that banks are *not allowed* to treat applicants differently based on race - given whatever features they use, they must make the same decision regardless of the race of the applicant. In the third regime (*No Disparate Impact*) banks must make loans at equal rates to each group, which implicitly *requires* incorporating race into their decision-making except for special cases.[9]

[8] That is, the COMPAS Northpointe controversy. See [1] and [58] for the initial arguments; [34] and [86] also provide summaries.

[9] For instance, if groups have identical feature distributions, then banks can lend at equal rates without imposing disparate treatment. But important features, like credit score, do vary by group; see, e.g., [107].

Since the principles behind both No Disparate Treatment and No Disparate Impact are desirable, law and policy tends to encourage both in various circumstances; hence it may well be that the real world exhibits neither Regime 2 nor Regime 3 but rather some combination of the two. These regimes are framed in absolutes (we have disparate treatment or not, we have disparate impact or not), but by viewing the *implications* of these regimes for various fairness definitions, which *can* be measured as continuous quantities, we can begin to think about these regimes as not all-or-nothing, but measurable shades of grey.

5.3.2 Theoretical Model

There is a representative bank. The bank will make a loan to a loan applicant with a single feature which we imagine to incorporate all information the bank has about an applicant's risk; we will call this feature X. Most naturally, we can think of X as a credit score[10], so we will refer to X as such. Applicants have a group membership $G \in \mathcal{G}$; for each group, the distribution of X is given by $\mathcal{D}_G$. We use d to denote whether the applicant defaults; in our context, we will assume that d is not deterministic in general, but rather a random variable, and that there is some function of X and G that maps the feature to the probability of default η. We will write the probability that $d = 1|X, g$ as $\eta(X, G)$, and we will write $\eta_G(X)$ for short. For every loan, if the loan is paid back, the bank earns return r, and if not, it loses cost c. That is, we make the simplifying assumption for this work that returns and costs do not depend on X or G. A *policy*, ℓ, is a map from applicant information (e.g. X, G) to a decision of approving or denying a loan, which we denote by 1 or 0 respectively. We will sometimes abuse notation to also write ℓ for the decision variable.

[10]Banks of course do have and will use more information than a single credit score, but the intent behind FICO and other credit scores is to serve as a summary measure capturing as many relevant factors for risk determination as possible.

Thus, for a given policy ℓ and realization d, the profit π of the bank is:

$$\pi(\ell, X, g) = \begin{cases} 0 & \ell(X, G) = 0 \\ r & \ell(X, G) = 1 \text{ and } d = 0 \\ -c & \ell(X, G) = 1 \text{ and } d = 1 \end{cases}$$

We assume that the bank is an expected profit maximizer, so its general problem is to maximize profit over the space of policies:

$$\max_{\ell} \mathrm{E}[\pi(\ell)] = \max_{\ell} \sum_{g \in \mathcal{G}} \Pr[g = G] \cdot \mathrm{E}[\pi(\ell, X, g) | g = G] \tag{5.3.1}$$

We will make the following assumption to simplify our thinking (though relaxing does not greatly affect what results are achievable):

Assumption 2. *$\eta_g(X)$ is monotonically decreasing in X.*

We restrict our attention to the class of *threshold* policies. In our setting, a threshold policy is a policy that tracks some feature or function of features (e.g. X or $\eta_G(X)$) and jumps from $\ell = 0$ to $\ell = 1$ at some point τ. Why should we limit ourselves to such policies? It is easy to see that among any deterministic[11] policy (i.e. any policy where our only choices are to make or not make a loan), any non-threshold policy cannot be optimal.

Regime 1: Unconstrained Profit Maximization Suppose that the bank has no constraints on what it may do. Then it is free to optimize its decision for each value of (X, G) separately. So rather than looking at equation 5.3.1, we can write its problem as:

$$\max_{\ell(X,g)} \mathrm{E}[\pi_\ell] = \max_{\ell(X,g)} \sum_{G \in \mathcal{G}} \Pr[g = G] \sum_{x} \mathrm{E}[\pi(1, X, g)] \cdot \mathbf{1}[\ell(X, g) = 1] \cdot \Pr[x | g = G]$$

[11] There is argument over whether randomized policies can be fair (see, e.g. [97]), and certainly they are not currently employed in high-stakes settings in practice.

And so we can simply consider the expected profit of making a loan conditional on (X, G), which is given by:

$$\mathrm{E}[\pi|\ell = 1, X, g)] = (1 - \eta_g(X))\, r - \eta_g(X)c$$

and the expected profit of not making a loan is, of course, 0. Thus, the profit-maximizing policy will satisfy:

$$\ell(X, g) = 1 \iff (1 - \eta_g(X))\, r - \eta_g(X)c \geq 0$$

Rearranging gives the following proposition:

Proposition 1. *The optimal unconstrained policy for the bank is:*

$$\ell(X, g) = 1 \iff \frac{1 - \eta_g(X)}{\eta_g(X)} \geq 0 \iff \eta_g(X) \leq \frac{r}{r + c}$$

In other words, the bank sets a single threshold on default probability – that is, two thresholds on credit score – that depends on the ratio of returns and costs and accepts an applicant if and only if their default probability below that threshold. If we define x_g^* as the minimum credit score for each group such that $\eta_g(x) \leq \tau$, then we see that the bank's policy will be to accept an applicant of group G if $X > x_g^*$. Since groups have different functions $\eta_g(X)$, this will result in a different threshold on X for each group.

Corollary 6. *Unconstrained, we have the following fairness properties in general:*

1. $Pr[\ell = 1|G] \neq Pr[\ell = 1|G']$ *(No Demographic Parity)*
2. $Pr[\ell = 1|X, G] \neq Pr[\ell = 1|X, G']$ *(No Conditional DP*[12]*)*

[12]One may also argue the equation given here is asking for No Disparate Treatment. We do not present it that way for two reasons. First, Disparate Treatment is conceptually about how decisions are made, i.e. starting at the beginning of the data generating process *looking forward*, while Conditional Demographic Parity is measured beginning with some dataset which is the result of a data generating process and attempting to infer whether that process was fair, and thus in a

3. $Pr[d = 1|X, G] \neq Pr[d = 1|X, G']$ *(No Predictive Parity on Average)*

4. $Pr[d = 1|X = x^*_G, G] = Pr[d = 1|X = x^*_{G'}, G']$ *(Predictive Parity at the Margin)*

Regime 2: No Disparate Treatment In the No Disparate Treatment regime, we assume that the bank is prohibited from treating individuals systematically differently based on their group membership; in this simple setting, we operationalize that constraint as requiring the bank to set a single threshold on X which it must apply to all groups. As the bank still wishes to maximize its profit, its problem will be to find the optimal *single threshold* on credit score which maximizes its profits.

Rewriting the general problem in equation 5.3.1, the bank's problem now is to maximize:

$$\max_{x^*} \mathrm{E}[\pi] = \max_{x^*} \sum_{G \in \mathcal{G}} \Pr[g = G] \, \mathrm{E}[\pi(\ell = 1, X, g)|X > x^*, g = G]$$

where we have simply replaced the full policy space with the set of credit score thresholds. Notice that the objective function must now incorporate not only each default map η_g, but also the underlying population share of each group. Consequently, the optimization problem is entangled in the sense that if we take the first order condition, the function for which we would search for the optimum is a weighted combination of the derivatives of each group, and so the optimal threshold will depend on the the population shares as well as the default functions.

It is not necessary to solve for the optimal threshold, however, to make the following pronouncement, which follows directly from the assumption that there is

sense *looking backward*. The latter better matches the situation of an outside observer analyzing data. The second is that the apparent mathematical equivalence of these two concepts requires that X be the full set of features under both conceptualizations. For instance, if a bank applies No Disparate Treatment to a set of features X', and we observe only a subset of features X, then we may observe a disparity in loan probability conditional on X that would disappear if we had the full feature set X'.

a single fixed x^* for all groups and that feature and default distributions may differ:

Corollary 7. *Under No Disparate Treatment, we have the following fairness properties in general:*

1. $Pr[\ell = 1|G] \neq Pr[\ell = 1|G']$ *(No Demographic Parity)*
2. $Pr[\ell = 1|X, G] = Pr[\ell = 1|X, G']$ *(Conditional Demographic Parity)*
3. $Pr[d = 1|X, G] \neq Pr[d = 1|X, G']$ *(No Predictive Parity on Average)*
4. $Pr[d = 1|X = x^*_G, G] \neq Pr[d = 1|X^*_{G'}, G']$ *(No Predictive Parity at the Margin)*

Regime 3: No Disparate Impact In the final regime, the bank must loan to both groups at equal rates. To do this, it will (except in very special cases) necessarily have to use *different* thresholds on credit scores. (As discussed above, it still makes sense to use thresholds - there is no reason to accept worse applicants in lieu of better ones - but now these credit thresholds will differ by group.) Equivalently, the bank is picking the *same* top quantile of each group, whatever that quantile may be.[13] Thus, bank's problem is now to solve:

$$\max_{r^*} \mathrm{E}[\pi] = \max_{r^*} \sum_{G \in \mathcal{G}} \Pr[g = G] \, \mathrm{E}[\pi(\ell = 1, X, g) | \Pr[X > x^*] = r, g = G]$$

Corollary 8. *Under No Disparate Impact, we have the following fairness properties in general:*

1. $Pr[\ell = 1|G] = Pr[\ell = 1|G']$ *(Demographic Parity)*
2. $Pr[\ell = 1|X, G] \neq Pr[\ell = 1|X, G']$ *(No Conditional Demographic Parity)*

[13] Arguably, one could view this as a form of equal treatment, conditioned on race. But this is probably not in the spirit of the law, and has been made illegal in the Civil Rights Act of 1991 [64].

	Unconstrained	No Disparate Treatment	No Disparate Impact
Equality of:			
Thresholds	✗	✓	✗
Risk Thresholds	✓	✗	✗
Approval Rates	✗	✗	✓

Table 5.1: Theoretical predictions of fairness metrics

3. $Pr[d = 1|X, G] \neq Pr[d = 1|X, G']$ *(No Predictive Parity on Average)*

4. $Pr[d = 1|X = x^*_G, G] \neq Pr[d = 1|X = x^*_{G'}, G']$ *(No Predictive Parity at the Margin)*

5.3.3 Identification and Marginal applicants

Our models may seem overly reductive, as banks certainly use more than a single feature.[14] However, credit score is certainly a dominant feature [23], so this reductive model likely captures an important facet of reality. But this simplification may be an issue if, for instance, decision-makers have access to some set of features $\tilde{X}$ not observed in our data that are predictive of default and correlated with race. A decision-maker that was implementing No Disparate Treatment on the *augmented* feature set $(X, \tilde{X})$ would appear to be discriminating with respect to even CDP, let alone DP and PPV.

Without experimental variation, the main solution to this problem is attempting to compare candidates that are all *on the margin* – that is, candidates essentially as likely to be denied as accepted. By definition, these candidates are all similar in

[14] One could imagine, of course, turning all information into a single perfectly predictive score (the default probability itself would suffice for instance, if it could be known). But such a measure certainly does not exist in practice, even if it *is* what FICO scores and related scores aim to be.

terms of creditworthiness, so we may reasonably take variation in a given fairness measure as evidence of discrimination. The question, though, is how to identify "marginal" candidates. In this work,[15] we will treat marginal applicants as those in a particular credit range – 620-660 – based on documented evidence that this range is explicitly flagged being worthy of special attention, and above and below are sure bets [23], which we explain further in Section 5.4.1.

5.4 Results

In this section, we focus on our empirical results. Table 5.3 summarizes the results in a simple table.

5.4.1 Data

We draw data from two sources:

Home Mortgage Disclosure Act (HMDA) HMDA is a dataset of loan applications, which is collected by law and made available in anonymized form to the public. Almost all loan applications aside from very rural areas are subject to HMDA and so included in this dataset. HMDA data, until recently, contains few and relatively general features, including the outcome of the loan (approved, denied, and so on), lien status, loan amount, applicant income, property type, geographic features like zip code, and so on. Crucially for fairness purposes, however, it contains the applicant race and ethnicity, which does not exist in many other datasets (including Black Knight McDash). Unfortunately, because HMDA did not, until

[15] In on going work, we are currently using a fuzzy matching approach to identify applications that plausibly represent borrowers applying to more than one bank for their loan. Those that have at least one of their applications approved and one denied can be seen as on the margin under the assumption that banks do a reasonably good job at evaluating risk and tend to agree in their evaluations.

very recently[16], contain detailed information on mortgage applicants, we can only observe rich applicant and loan features via other datasets.

Black Knight McDash (McDash) Second, we use a proprietary anonymized dataset from **Black Knight McDash** (McDash), which tracks the performance of loans over time, and also more detailed information about the loan upon origination (the original FICO score of the applicant, for instance, the initial rate of the loan, the term, etc.). McDash covers a large fraction of loans as well, but importantly, loans in McDash are (by definition) originated loans, so we cannot use this richer data to better understand *denied* loan applications. We do not use McDash directly, however - instead, we use a unique matched dataset that joins McDash with HMDA data, because McDash does not contain race and so cannot be used alone for fairness purposes. This match is not exact as there is no shared identifier across the dataset, instead using loan characteristics, time, and geography to identify candidate matches. We limit our analysis to mortgages where there is a unique potential match candidate; this does not cover all of the loans in HMDA, but rather about 60-70 percent of them depending on the year. For computational reasons, we use a 10% sample of the overall HMDA-McDash match (sampled at the loan level).

Race In this work, we focus on four major groups which we denote by the term *race*. These are Asians, Blacks, Hispanics (that is, Hispanic Whites), and Whites (Non-Hispanic Whites). This is a reasonable choice – over the time period we work with, HMDA provides for each borrower racial categories that include Black, White, Asian, Hawaiian/Pacific Islander, and American Indian, (the latter two of which are very small as a fraction of the data) and ethnic status of Hispanic/Latino or Not Hispanic/Latino. The choice we makes allows us to avoid overlapping subgroups and focus only on groups large enough to be substantive in our 10% sample.

However, we do not mean to posit that these categories are the "right" notion

[16]See future work.

of race, or that any notion of "race" is right or correct; certainly, these groupings are far too coarse to capture the variety of individual identities that we might think of as race, ethnicity, culture, or other relevant groupings. Instead, we are simply attempting to use these classifications available to us in HMDA (which are based on notions of race used in the United States Census) to capture a notion of shared socioeconomic experience in keeping with the history of discrimination in the United States. Indeed, the Census itself has shifted possible options for race several times in history; in more recent iterations it allows for much finer notions of race and ethnicity. It would certainly be interesting to study fairness among these finer or overlapping notions, though it is also worth keeping in mind that finer group notions come at the cost of smaller sample size and precision.

FICO and LTV For most of our empirical work, we will focus on two features which are available in McDash: Loan-to-Value ratio (LTV), and FICO score (FICO). These features are known to be crucial in the determination of creditworthiness according to both lenders and GSES [3]. LTV is simply the size of the loan relative to the total value of the property (at the time of origination) - so for instance, an applicant applying for a mortgage with a 20% down payment would be applying with an LTV of 80%. The FICO score, pioneered by the Fair Isaac Corporation, combines various pieces of information in individuals' credit history to come to a general measure for riskiness.

The FICO score is extremely widely used, and in particular, was adopted by GSEs, as discussed in Section 5.1.1, as tool for evaluating credit. Indeed, because of the market power of the GSEs as the largest buyer of mortgages from private originators [3] to set credit threshold for loans they would or would not buy, their decisions had a huge impact on the market.

Indeed,[23] provides a description of the thresholds suggested by GSEs:

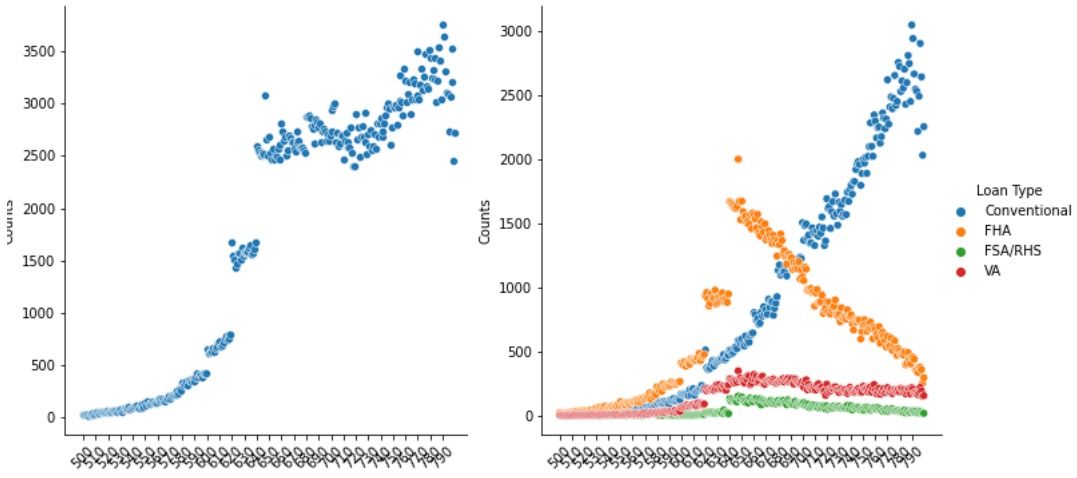
> [23] In 1995 Freddie Mac sent a letter to originators directing them to

> begin using credit scores in underwriting and establishing three tiers of credit scores (Freddie Mac, 1995). We provide the key part of the letter in the supplemental appendix. The FICO scores of 620 and 660 were important cutoffs. For borrowers with FICO scores above 660, lenders were to do a "basic" review to "underwrite the file as required to confirm the borrower's willingness to repay as agreed." For borrowers with FICO scores between 660 and 620, lenders were to perform a "comprehensive" review to "underwrite all aspects of the borrower's credit history to establish the borrower's willingness to repay as agreed." For borrowers with FICO scores 9 below 620, lenders were warned to be "cautious" and to "perform a particularly detailed review of all aspects of the borrower's credit history to ensure that you have satisfactorily established the borrower's willingness to repay as agreed." Fannie Mae (1997, pp. 8–9) established a similar set of cutoffs, including at both 620 and 660 FICO. Lenders who sold loans to Fannie Mae and Freddie Mac were contractually obligated to follow the GSEs' guidance letters establishing credit score cutoff rules for screening.

A consequence of these threshold-setting decisions is that the applicants between 620-660 are borderline, while those above 660 are very likely to be accepted and those below 620 very likely to be rejected. The category of 620-660, then, should contain those candidates that are most likely to have some positive probability of being approved and of being denied. We will thus treat these candidates as "marginal" even if they may not be exclusively marginal in the sense of being as likely to be denied as approved (and there may be other candidates who are deemed marginal based on other features).

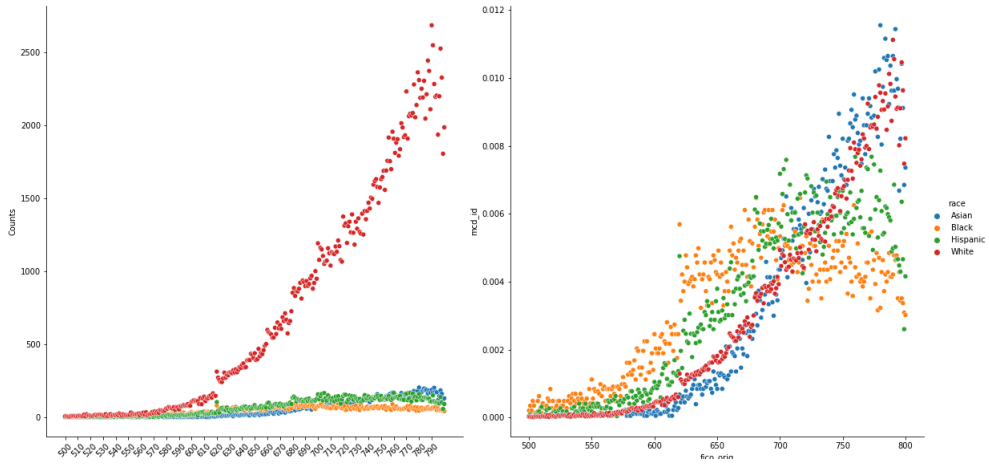
We can observe the consequences of these thresholds empirically. Figure 5.2a displays the number of mortgages in our sample at each FICO bucket, while Figure 5.2b breaks it down by loan type. Notice that there are discrete jumps at thresholds

in buckets of 20 points, and indeed, the largest seems to be at 620.



(a) Counts by FICO (b) Counts by FICO by Loan Type

Figure 5.2: FICO scores of mortgage holders at time of origination . Source: HMDA-McDash



(a) Count of FICO at origination among mortgages by race (b) Fraction of borrowers at each FICO at origination by race.

Figure 5.3: FICO by race among borrowers . Source: HMDA-McDash.

Table 5.2 provides an estimate of the population-wide FICO distribution according to FICO's own blog [74]) as of October 2012, near the middle of our time

FICO range	300-499	500-549	550-599	600-649	650-699	700-749	750-799	800-850
Percentage Points:								
Population-wide ([74])	6.0	8.5	9.9	10.1	12.2	16.2	18.8	18.4
Overall Borrower Sample	0.1	0.6	2.4	13.0	24.3	24.7	28.9	6.2
Conventional-Only Sample	0.1	0.4	1.2	5.9	15.4	26.9	40.7	9.3

Table 5.2: Population-wide FICO distributions. Source: HMDA-McDash and [74].

sample, as well as our observed FICO frequency among our sample of borrowers. These differ for several reasons. First, of course, we observe a much higher proportion of borrowers at higher FICO scores and lower proportion at lower scores relative to the population-wide estimate, which makes sense, since low-FICO potential borrowers are far less likely to receive (or perhaps even apply for) a loan. Second, we are focusing on 30-year fixed rate mortgages that are owner-occupied, and different groups may prefer different products; this may explain, why, for instance, our sample has a lower share of very high FICO score borrowers than the population as a whole. Finally, it is instructive to note that conventional loans have a much higher proportion of higher borrowers relative both to the population as a whole and to all successful borrowers.

5.4.2 Credit Score Threshold

The first prediction we test is that of a uniform credit score threshold. Recall that in our simple model, both the unconstrained bank and the bank under No Disparate Impact will choose different thresholds[17] for each group, while the bank under No Disparate Treatment will chose the same threshold.

Notice that the discussion in Section 5.4.1 strongly suggests that there ought to be a (soft) threshold. How would we identify such a threshold in our data? Again, because we do not have FICO or LTV of denied applicants, we cannot directly

[17]We can say more about the relationship between the different thresholds if we make further assumptions, such as the $\eta_g(X)$ functions being non-crossing, but not in general.

identify a threshold for approval. However, we can look at the distribution of FICO and LTV scores among originated mortgages and observe whether the distribution of FICO scores spans the whole range of possible scores. Figure 5.2a shows borrowers at each FICO score, and there is a large jump around the purported threshold of 620. This jump might not be as large as expected, but as visible in Figure 5.2b, there are very few conventional mortgages below a FICO score of 620 - instead, other loans like FHA loans make up the bulk of low-score loans.

But as noted, FICO is not the only factor – a higher FICO score might be required at a higher LTV, for instance. Figure 5.4 is a heatmap of originated mortgages by FICO at origination (grouped into buckets of 20 points) and LTV (grouped into buckets of 5 percentage points), while Figure 5.5 shows the same quantities but limited to each race. Notice that in the overall plot, the shading drops very quickly below 620, suggesting that this is indeed a threshold. However, breaking this figure down by race shows a more nuanced picture. For Whites and Asians, these 620 threshold seems to hold, with a few observations below suggesting some wiggle room. On the other hand, for Blacks and Hispanics, there is a substantial number of observations below the 620 purported threshold. These observations tend to be near 95% LTV –i.e. with *less* downpayment than the more standard 80%[18] – which does not concord with a possible explanation that the threshold is relaxed for higher down payments. Instead, this may be consistent with attempts to ameliorate disparate impact, perhaps consistent with policy goals like the Community Reinvestment Act.

[18] An 80% loan-to-value ratio corresponds to a 20% down payment.

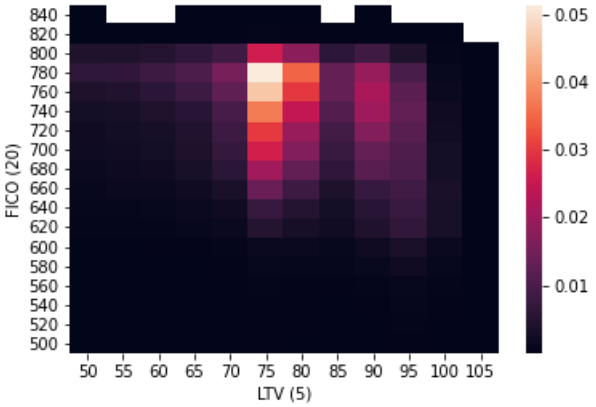
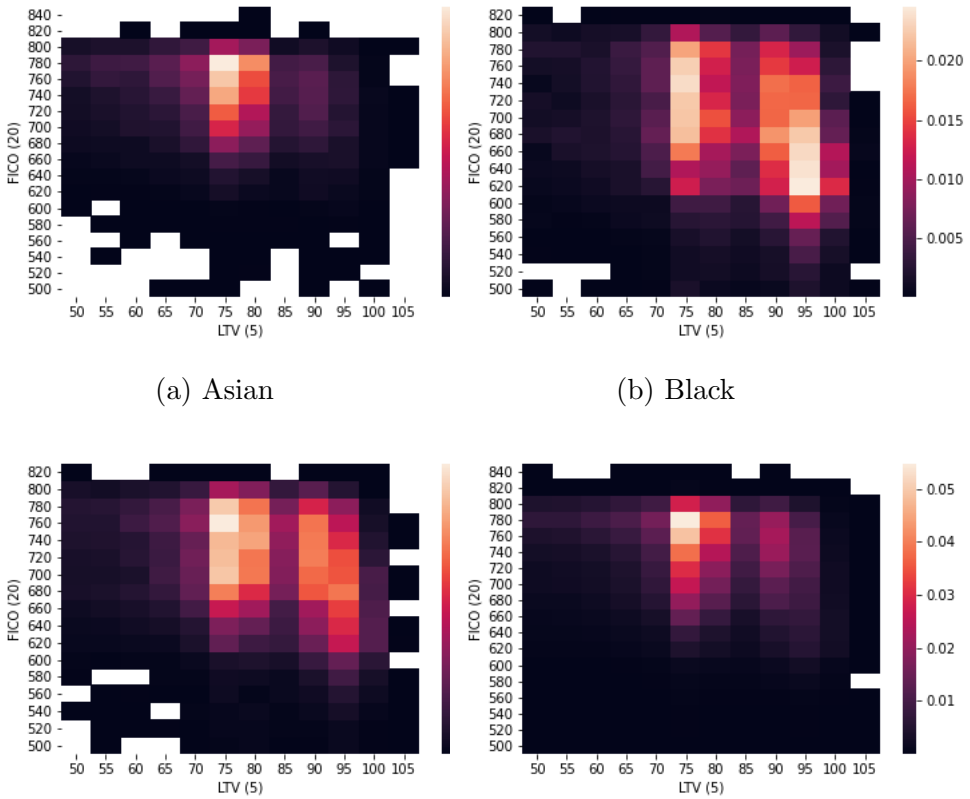


Figure 5.4: Originated Mortgage LTV-FICO heatmap for conventional loans . Source: HMDA-McDash

	Asian	Black	Hispanic	White	Overall
Quantity:					
Fraction loans with <= 620 score	1.0%	13.4%	6.4%	2.4%	3.2%
Default Probability at Margin	17.2%	27.6%	25.6%	16.7%	19.5%
Overall Default Probability	3.2%	14.4%	11.4%	3.4%	4.5%
Denial Rate	14.4%	28.9%	22.3%	11.6%	15.0%

Table 5.3: Identified quantities by race (Conventional loans only) Source: Source: HMDA; HMDA-McDash



(a) Asian (b) Black

(c) Hispanic (d) White

Figure 5.5: Originated Mortgage LTV-FICO heatmap by race Source: HMDA-McDash.

5.4.3 Risk of Default

The next prediction we turn to is that of the default risk. Recall that an unconstrained profit maximizing bank would set differing thresholds, but these thresholds would have *equal* default rate risks.[19] On the other hand, banks that must avoid

[19] An objection is that perhaps the loans chosen by some groups are different than others in a way that affects their profitability; this could result in differing risk thresholds even by a profit-maximizing bank. This is an important area for future work and highlights the complexity of measuring fairness this in setting. Importantly though, some natural straightforward extensions for realism will not affect this result – for instance, merely differing scale applied equally to rewards

disparate impact or disparate treatment would in general have different default rates on the margin and overall.

We thus turn to those applicants around the credit score threshold - these marginal candidates- and examine their default risk. For our purposes, we will define a loan in default if it becomes 90 days delinquent (according to the Mortgage Bankers Association method of measuring days since payment due date) within the first three years. This restriction to the first several years is a practical one, but is common in the literature (e.g. [60]) as it allows us to evaluate mortgage outcomes without waiting the full lifetime of the loan, and helps us avoid comparing apples-to-oranges in the sense of loans of different lengths.

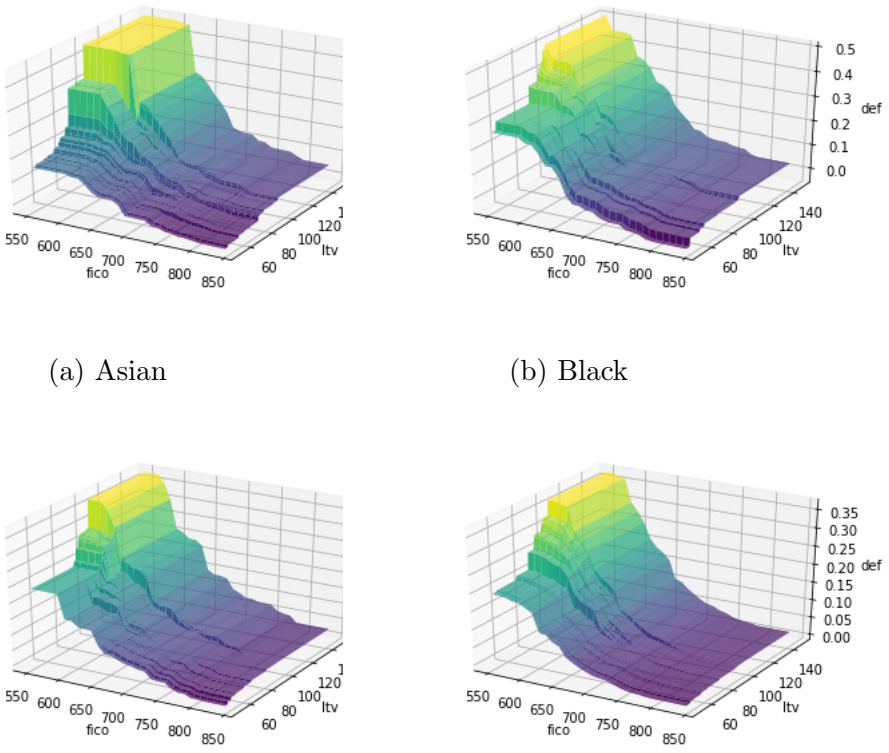
Note that for this exercise, we would like to make aggregate default *predictions* as accurately as possible. To do so, we estimate a histogram-based gradient-boosted regression tree (constrained to be monotonic) on FICO and LTV of our borrower dataset. Because we are focusing solely on FICO and LTV, a model is not strictly necessary; for example, we could use a histogram approach on the empirical data. But such an approach imposes assumptions – e.g. how fine the appropriate resolution of grid size is, whether that grid should be uniform and if not, how should it vary, and so on. Instead, we can use a prediction model to implicitly learn what values should be predicted for each combination of FICO and LTV in service of some goal, i.e. minimizing square error. We choose a tree-based algorithm for several reasons: first, such algorithms can more easily fit the possibly complex and nonlinear patters that may appear, and have been shown to succeed empirically in state-of-the-art prediction results [81], [66]. By contrast, algorithms like logistic regression impose linearity, which makes them interpretable and useful for inference but poorer for inference [69]. Finally, tree algorithms are generally memory-efficient

and costs will not affect the equal thresholds results. Additionally, we highlight that in perfect competition, profits ought to be competed away (in our model, that would mean that each group would have a marginal default rate of $r/(r+c)$) so the more competitive we are, the more this equal risk threshold analysis would be expected to hold.

and can be trained easily on large datasets.

To train our models, we use the histogram-based gradient-boosted regression tree (HGBRT) from Sci-kit Learn [103]. For each of the following subsets – Asian, Black, Hispanic, White, and all borrowers – we train separate models, and create train-test splits with a 20% test size. The outcome is a binary variable representing defaulting within the first three years or not, and the only variables we use are FICO and LTV at origination. We also impose a monotonicity constraint in both dimensions (positive in LTV, negative in FICO) in order to regularize the models and also impose our prior belief that these variables should be at least somewhat predictive of risk.

We focus on the separately trained by race models because they allow for better accuracy. The prediction surfaces we arrive at are given in Figure 5.6. We note that for each group, the general trend is as expected – high LTV and low FICO have very high risk of default, while low LTV and high FICO have very low risk of default – and the effect of FICO appears much stronger than LTV. Moreover, the steepness of the relationship between default risk and FICO seems to interact with LTV, with LTVs near and above 100 having much steeper declines than at lower LTVs. Overall, the shape of the curves are highly non-linear.



(a) Asian

(b) Black

(c) Hispanic

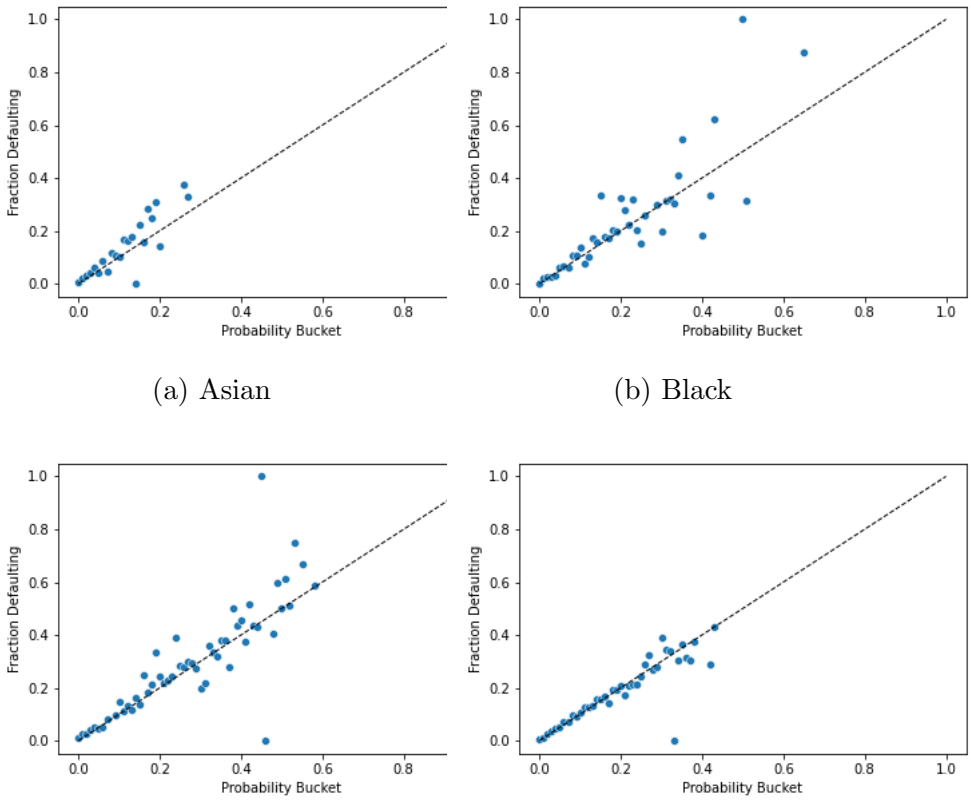
(d) White

Figure 5.6: Learned default prediction surfaces. Source: Author calculations using HMDA-McDash.

So the models are intuitive. But how accurate are they? It is in general not as easy to score models that predict probabilities rather than classes, since there is random noise in the outcomes of realizations, and moreover, economic settings in particular face the likelihood that random shocks or market-wide changes can greatly affect individual outcomes. So rather than thresholding at a particular value, we prefer to test whether the models are *calibrated*[20] – that is, of borrowers for whom our model predicts a default probability of k, does about a k-fraction of these borrowers default? The results are given in Figure 5.7. We divide our model's

[20]Note that this is another fairness metric, but not one we examine in detail here.

predictions into probability buckets of 1 percentage points – that is, borrowers with a 1% predicted default chance, a 2% predicted default chance, and so on – and evaluate the realized default rate in the *test* dataset. A perfectly calibrated model would be exactly on the line $y = x$, which is plotted as a dashed line; we see that each of the points is very nearly there, with some deviations. By comparison, training a single HGBRT on all borrowers gives a somewhat noisier fit, in some cases systematically. We display test set calibration plots for that model in Figure 5.20. On the other hand, training separate *logistic regression* models to predict default and following a similar approach results in the calibration plots in Figure 5.21.



(a) Asian

(b) Black

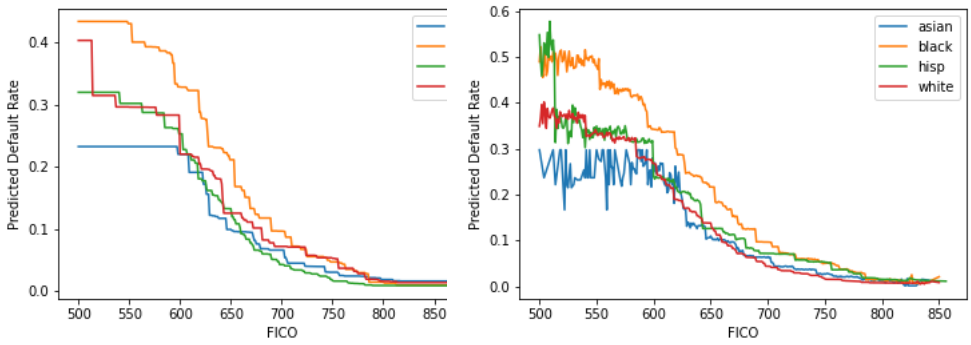
(c) Hispanic

(d) White

Figure 5.7: Default model calibration on test set (models trained separately by race) Source: Author calculation using HMDA-McDash.

With the HGBRTs we trained, we can now estimate a race-specific default probability for each race at each FICO score. We limit to borrowers between 80-100 LTV, since these are likely most representative of the standard setting. We plot these estimated default probabilities in Figure 5.8, under two assumptions on LTV: in 5.8a, we average predictions at a given FICO score uniformly over an LTV of 80-100; in 5.8b, we instead predict over the LTV distribution at each FICO score observed in the data. (The former, while less natural, holds equal the LTV distribution across groups, and so is useful for eliminating the composition effects of different LTV ratios; on the other hand, the latter better captures observed defaults in practice.)

This gives the results in the second row of Table 5.3: the overall marginal default rate is about 15%, with the estimated default probability being 11.9% for Whites, 14.7% for Asians, 20.6% for Hispanics and 25.6% for Blacks.



(a) Uniform LTV. (b) Empirical FICO/LTV.

Figure 5.8: Model-estimated default probability by FICO and race. Left: Calculated over uniform LTV 80-100. Right: calculated according to empirical FICO/LTV distribution. Source: Author calculation using HMDA-McDash.

Overall, these default rates are significantly different at the margin – Hispanics and Black on the margin default at nearly or more than double the probability of Whites on the same margin. This difference is again consistent with an attempt to mitigate disparate impact.

5.4.4 Demographic Parity

Finally, we turn to demographic parity. For this section we again focus on conventional loans. As described in Table 5.3, Blacks are more than twice as likely to be denied for a loan as Whites – 25.6% vs 11.9% – while Hispanics(20.6%) are nearly twice as likely to be denied and Asians (14.7%) are 1.25 times as likely to be denied. Recall that we expect an extreme version of the No Disparate Impact regime to have equal denial rates, while unconstrained or No Disparate Treatment

regimes will tend to have disparate denial rates. The fact that these rates are so disparate suggest that we are certainly not in a full No Disparate Impact regime[21].

What about conditional demographic parity? First, Figures 5.9 and 5.10 display the difference in Black vs. White denial rates in each state and county, respectively. There appears to be a large amount of variation - some states, such of those of the upper Midwest and south, seem to have substantial differences, as much as 20 percentage points[22], while others, such as in the West, have much less disparate impact. However, it is difficult to interpret these figures too much – there are a multitude of historical and incidental factors that may generate these results. There may also be mundane factors like sample size that limit the interpretability of these figures – for instance, Hawaii and Montana have some of the lowest differences (just 0.1% and 3.9%, respectively), but these states have very small black populations (about 2100 and 4000, respectively) as of the 2010 U.S. Census.

Evidently, understanding conditional demographic parity requires a fuller analysis. But it is instructive to note that two naive regression approaches, with results detailed in Tables 5.7 and 5.8 (OLS and Logistic regression, respectively), do not appear to explain away the differing approval rates. These regressions control for state and year fixed effects, presence of a co-applicant, and loan-to-income ratio, and retain relatively large coefficients on dummy variables for race. Unfortunately, because HMDA does not provide FICO or LTV until very recently, we cannot include the controls that would likely be the most important. However, as HMDA has added these beginning in 2018, we will engage in future work that can at least measure CDP with these two features going forward.

[21] Which is consistent with the fact that even the Disparate Impact legal theories of discrimination recognize business-necessity.

[22] In Michigan, the denial rate was 34.3% for Blacks and 13.8% for Whites.

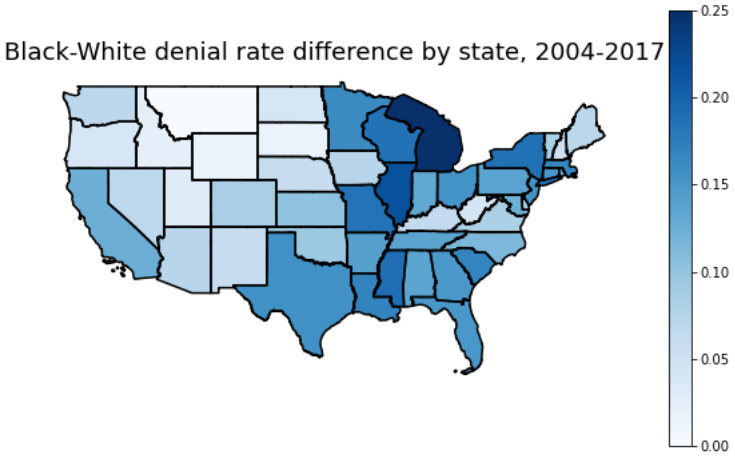


Figure 5.9: Black-White denial rate difference by state over 2004-2017. Source: HMDA

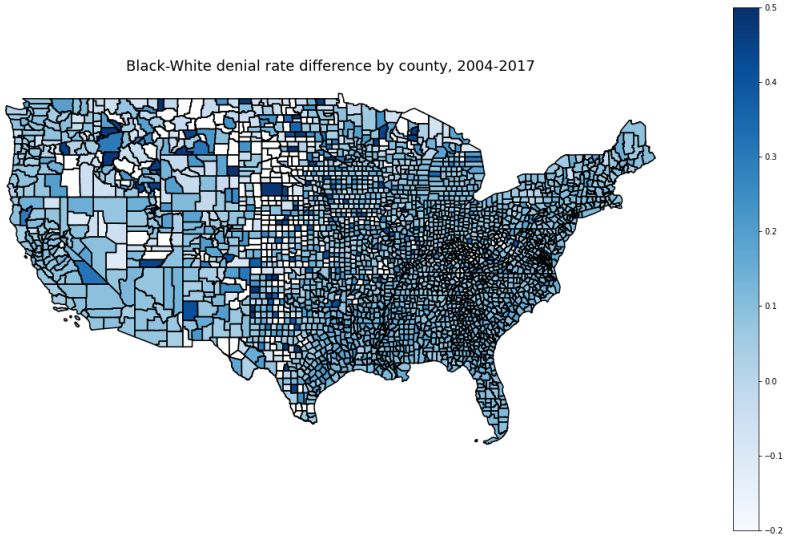


Figure 5.10: Black-White denial rate difference by county over 2004-2017. Source: HMDA

5.4.5 Summary

Sections 5.4.2 - 5.4.4 evaluate the various fairness metrics described in Section 5.2.1 for which our theoretical models make predictions. Overall, our evidence is mixed. We do see a threshold-like pattern around a 620 credit score, consistent with stylized facts about lender incentives, but the threshold appears to be less strictly enforced for minorities. We do not see equal default rates at among marginal candidates of different groups – instead, minorities tend to have higher default rates even at the margin. And we do not see similar approval rates across groups – instead, minority applicants are denied significantly more often. Taken together, this mixed evidence suggests that we are not in either a strict No Disparate Treatment or No Disparate Impact regime, but somewhere in between. Since the principles behind both regimes seem to be enshrined in the law, this may not be so surprising, but we have yet to formalize a notion of an "in-between" or hybrid regime. In the next section we will do so, and identify the trade-offs under such a regime or strict No Disparate Impact or No Disparate Treatment regimes.

5.5 A Counterfactual Pareto Frontier

Sections 5.4.2-5.4.4 focus on measuring fairness in the world as it is. Now, we want to consider the space of what could be, so that we can understand what choice we, as a society, are implicitly making. To do this, we consider possible trade-offs under various policy regimes and construct *counterfactual* Pareto frontiers. The three regimes we consider are No Disparate Treatment, No Disparate Impact, and a "decoupled" regime of separate policies on credit score. Given our data, we will have to make significant assumptions to do so; so as will become clear, the results in this section should be considered strictly proof of concept. In future work, we hope to obtain more detailed data and identify quasi-experimental variation; this will allow us to estimate similar quantities, but with greater quantitative credibility. In

particular, an important task we must do (which could be answered with existent, if proprietary, data) is estimating the *group-specific* FICO distributions.

Now we briefly introduce some notation for this section to clarify the exposition. Because we are working in a specific setting, we will shift away from writing X for a generic feature instead write F to emphasize that it is the FICO score in particular. We will use $\alpha_\Pi(G)$ to denote that approval rate of group G under policy Π – that is, $\Pr_{f\sim\mathcal{D}_G}[\ell_\Pi(f,g)=1|g=G]$ – but will omit the Π when it is clear. For threshold policies on credit score, we will again use τ, and write τ_G if we mean to allow thresholds to differ by race. We will use θ for thresholds on percentiles relative to the population, and to specify relative to the group population we will write θ_G. We write $\delta_G(F)$ as the default probability $\Pr[d=1|f=F,g=G]$. We will generally use the $\hat{\cdot}$ symbol to denote estimated quantities, either empirically or using a model. For instance, we write $\hat{\delta}$ for our estimate of a default rate. We will denote models, e.g. for prediction, with Θ; for instance, $\hat{\Theta}_G^{\text{HGBRT}}$ is the HGBRT we trained for group G in Section 5.4.

Throughout this section, we will have to make the following two assumptions:

Assumption 3. *We assume that* $Pr[d=1|F=f,g=G,\ell=1]=Pr[d=1|F=f,g=G,\ell=0]$, *where* ℓ *is the policy corresponding to the real world. That is, observed conditional default rates in McDash correspond to default rates for similar individuals who did not receive a loan. (We will thus write* $Pr[d=1|F=f,g=G]$*).*

Assumption 4. *All distributions are independent of our chosen threshold. (In other words, the* act *of changing our threshold will not invalidate the other assumptions.)*

Assumption 3 is necessary because we only have default rates for borrowers who obtained a loan; without this assumption, we cannot say anything about the default rates of borrowers below the threshold of loans made. Assumption 4, like Assumption 3, is quite important. Without it, it is in principle possible that the

act of changing our policy could greatly impact the behavior of consumers and underlying default rates.

These assumptions, while important caveats to our counterfactual approach, do not entirely eliminate its utility for several reasons. First, even if we cannot claim that we know the default rates of borrowers who would not have been made a loan under the historical policy, we at least have high confidence in the sign – selection bias will likely be only causing us to underestimate the default risk under the threshold (since borrowers who are *even more* likely to default would be unlikely to receive a loan and thus be in our sample). This allows us to at least view our results as lower bounds. But more broadly, we can ultimately estimate the default rates among those not offered a loan if we can either conduct experiments or identify quasi-exogenous variation ("natural" experiments) in the data. While it is hard to be systematic about identifying and collating natural experiments, natural experiments tend to be numerous with enough care and effort. As for Assumption 4 – first, it is almost certainly true that bank policies can create feedback loops and change incentives that end up changing behavior in the long run. But given how costly loans are, and how difficult it is to change features of one's creditworthiness, it is likely that this assumption will hold approximately, at least in the short tun.

5.5.1 Estimating the credit score distribution

The first task we need to achieve is to estimate the population-wide distribution of credit score by group. As discussed in Section 5.4.1, we cannot simply use the distribution of credit scores in our HMDA-McDash sample, because those who received mortgages (who are the only potential borrowers whose credit scores we have) are likely to have higher credit scores than those who did not receive or did not apply for mortgages. Unfortunately, publicly available detailed data on the credit score distribution by race is exceedingly rare [107].

To mitigate this issue, we will use one of the most detailed estimates of credit

score distribution by race – which comes from a study conducted by the Federal Reserve Board of Governors and reported to Congress in 2007 [107]. As part of an examination of possible disparate impact in credit scoring, the authors combined credit data with Social Security Administration data, which includes demographics, and estimated the fraction of each race that fell into the ten deciles of overall credit scores (scores were normalized for compatibility). The credit scores they used were provided by TransUnion, a consumer credit reporting agency; the scores provided were the Transrisk Account Management Score (Transrisk score) and the Vantagescore. The Transrisk score was developed by Transrisk, and the Vantagescore was developed in a joint venture between TransUnion, Equifax, and Experian in an attempt to harmonize scoring across their separate agencies. Hence, these scores are not exactly the FICO scores, but they are on the same scale of 300-850[23] and, though exact details are not available publicly, are generally believed[24] to be intended to track FICO scores. The report notes that their results across both scores are nearly identical.

While likely slightly out of date, and not technically based on the same score, we expect that these race-specific credit distributions will provide a plausible proxy for the shape of FICO. We will thus use these figures, in combination with the population-wide FICO distribution reported in Table 5.2, to create an estimated race-wide FICO distribution. Our approach will be to use the (bucket-level) population-wide FICO distribution to recover FICO scores for each decile; we then assume that fraction of each race falling in these buckets correspond to those falling in the rank buckets in the Fed's report. The implied decile FICO buckets are somewhat irregular in size, however, which presents a problem – it is clear that the distribution of scores is nonlinear, but it is not clear how scores are distributed within buckets. The approach we take is to assume that scores are distributed uniformly *within*

[23] See https://www.transunion.com/resources/indirect/doc/products/resources/product-creditscore-comparison-chart.pdf

[24] See https://www.doctorofcredit.com/credit-scores/fako-score/transrisk-score/

buckets. Because we still have irregular buckets overall, we maintain a global non-linearity despite accepting a local linearity that is probably not reflective of the true distribution.

Now we describe the steps in more detail:

Step 1: Obtain FICO rank distribution. Because the Federal Reserve's report intended to compare several credit score models, including their own, the authors of that report chose to normalize each of the scores into a rank-ordered scale. Table 5.4 reports the Transunion FICO scores in a nationally representative random sample; unfortunately, the authors only provided these normalized scores, rather than scores by FICO bucket, and do not provide a mapping from these scores to the actual FICO scores. We will, however utilize these figures to obtain a plausible estimate of race-specific FICO score distributions in Steps 2 and 3.

Notice that the distribution over buckets is quite different across races – for instance, 30 % of Blacks were in the lowest decile, compared with almost 6% of Asians, 15% of Hispanics, and almost 8% of Whites. Conversely, the highest decile contains 9.8% of the Asian population and 12% of the White population.

Decile	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Share of Race										
Asian	5.7	6.6	7.3	10.6	12.0	14.0	12.4	11.1	10.6	9.8
Black	30.1	22.5	15.6	10.1	7.2	4.6	3.2	2.7	2.4	1.7
Hispanic	15.1	15.0	14.9	13.3	10.8	9.5	6.8	5.6	5.0	4.0
White	7.8	8.5	8.7	9.9	10.1	10.0	9.7	10.6	12.7	12.0

Table 5.4: Estimated share over deciles of credit score by race, according to [107]
Source: Federal Reserve Report To Congress

Step 2: Construct population-wide credit score deciles. Using [74]'s population-wide FICO distribution as given in Table 5.2, we next construct an estimate of credit score deciles (which we will need because Table 5.4 only reports shares in terms of where individuals fall over the deciles). Because these figures are very high-level, we will need to interpolate them. In particular, we will interpolate linearly within buckets, and join or split buckets where necessary. For instance, the population estimate of the proportion of borrowers with FICO between 300 and 499 is given as 6%, and the proportion of borrowers with scores between 500-549 is given as 8.5%. Assuming that mass is distributed linearly within-buckets, the 10th percentile would be reached at a score of $499 + 4/8.5 * 50 \approx 523$. The next decile would start at 523, with the remaining 4.5% of the 500-549 bucket forming the beginning of the second decile, and extend to $549 + 5.5/9.9 * 50 \approx 577$. If we continue this procedure, we obtain the following table:

Population Decile	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Upper Boundary of Decile	523	577	626	670	707	737	764	790	817	844

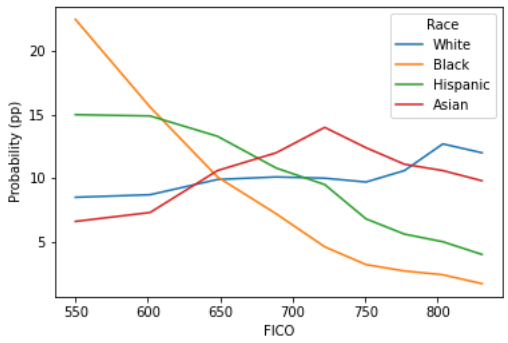
Table 5.5: Estimated share over deciles of credit score by race, obtained by combining [107] and [74]. Source: Author calculation using [107] and [74].

Notice that even in the population as a whole, the size of deciles varies widely – the first decile stretchs from the minimum theoretical score, 300, all the way to 523, while the ninth decile spans just 790-817. This global nonlinearity is an important characteristic of the distribution.

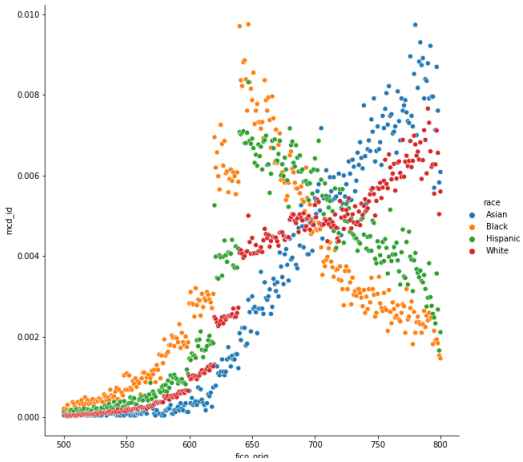
Step 3: Allocate share of race across appropriate deciles. Now, we construct each race-specific FICO distribution by allocating the share of each race in a given rank decile uniformly across the associated credit score bucket. For instance, our estimated bucket range for the third decile is 577-626, and the share of each

race that falls into the third decile for Asians, Blacks, Hispanics, and Whites is 7.3%, 15.6%, 14.9%, and 8.7%, respectively; this approach would then posit that 7.3% of Asians are uniformly spread over 577-626, 15.6% of Blacks, and so on. Since there are about 50 FICO points in this range, we would place roughly 0.37%, 0.78%, 0.75%, 0.44% of Asians, Blacks, Hispanics, and Whites on 577, the same number on 578, on 579, and so on. While this decision to spread the borrowers out uniformly does not add information relative to the mean value (by definition), it does allow us to consider smaller changes at a time, and the approximation seems reasonable.

Figures 5.11-5.12 display our estimates of the probability mass and cumulative density functions. There are several key points to note. First, our estimates largely concord with result in the literature that suggest (in less detail than we need) that Blacks and Hispanics have a significantly and systematically lower credit score than Whites and Asians. Second, this difference is not merely a shift in averages, but one of shape – Asians and Whites tend to have a uniform or slightly increasing probability mass function (PMF) overall, while Blacks and Hispanics have decreasing ones, suggesting that significant compositional shift may be observed when shifting thresholds. Finally, the probability distribution estimated differs greatly from the observed distribution of credit scores in HMDA-McDash (displayed in Figure 5.11b). For instance, the share of borrowers below 620 is far larger than observed in the empirical distribution among mortgage holders; this is expected, but indicates that we likely would not achieve plausible counterfactuals without a process of estimating the race-specific credit score distribution such as this one.

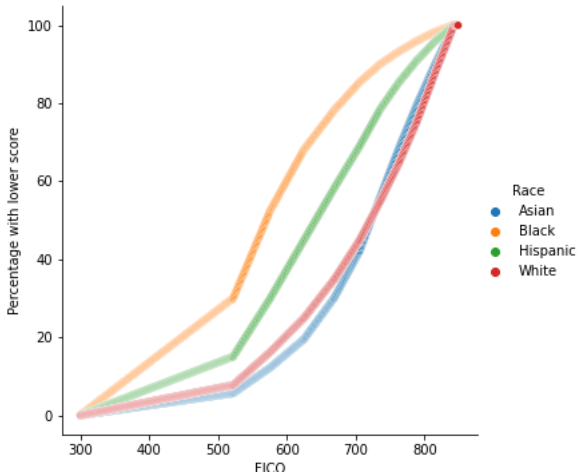


(a) Estimated population-wide FICO probability mass function (smoothed) Source: Author calculation using HMDA-McDash and results of calculations using [107] and [74].



(b) Frequency of Credit Scores among Borrowers. Source HMDA-McDash.

Figure 5.11: Population-wide estimates and observed frequency of FICO in HMDA-McDash.



(a) Cumulative Distribution Function

Figure 5.12: Estimated population-wide FICO cumulative distribution by race. Source: Author calculation using HMDA-McDash and results of calculations using [107] and [74].

Again, it may be that the true distribution within buckets may be far from linear (and we do not have enough information to gauge how far this assumption may be from reality); however, this approach preserves the *global* nonlinearity of the distributions, if not local, and spreading probability mass uniformly in the bucket is preferable to other presumptive approaches like placing all borrowers at the midpoint (which results in the same mean per bucket, but no variance), trying to estimate a distribution within buckets from our data (which would suffer from selection bias), or choosing some other unjustified distribution. Of course, we can sidestep all these estimation issues if we can obtain population-wide FICO distributions from the agencies; however, as they do not collect race by law, we would then face the alternate and also difficult task of obtaining detailed external data with demographics that could be matched to credit scores (and convincing the parties involved to allow the match).

Estimating Denial and Default Rates Now that we have credit score distributions, we can begin to ask what we can expect in terms of denial rates and default rates as we vary the threshold, whether expressed in terms of a FICO cutoff or a top percentile. In particular, we can easily read off groupwide denial rates at any given FICO cutoff from these estimates, and construct weighted default rates for each group using the probabilities above a threshold as weights and normalizing. And starting with a percentile, we can use the estimates to recover the corresponding FICO cutoff.

Thus, we have all the ingredients to estimate counterfactuals in strict No Disparate Treatment and No Disparate Impact regimes. It is worth noting, however, that the implied acceptance thresholds predicted in our estimates are not consistent with a *strict* No Disparate Treatment regime for Blacks and Hispanics. That is, given the documented evidence of a real-world cutoff at 620, we might expect that the fraction of applicants accepted would approximately match our estimates for the share of the population above the threshold. Yet for Blacks for instance, our estimate of the fraction of the population above the 620 cutoff is 33.8%, while the approval rate is 71.1% – a 37 percentage point gap. There are two reason why these might not match, besides our estimate being inaccurate. One is that the population of mortgage applicants is very different from the overall population in terms of credit score, so that approximately 71.1% of Black mortgage applicants do have a credit score above 620. This *selection bias* may be a large part of the explanation, since borrowers of very low credit score likely do not even apply for a mortgage knowing they will be denied. The other is the possibility that the thresholds on FICO are not being strictly applied for Black and Hispanic borrowers – perhaps because FICO scores are not as predictive for minority borrowers, or because banks wish to avoid disparate impact – and this is also consistent with the fact that we see a nontrivial fraction of minority mortgageholders with scores below 620.

We will thus proceed to estimate three sorts of counterfactuals. We begin with

Group	Asian	Black	Hispanic	White
Approval Rate	85.6%	71.1%	77.7%	88.4%
Share at least 620	81.4%	33.8%	56.8%	76.1%
Implied gap (percentage points)	4.2	37.3	20.9	12.3

Table 5.6: Estimated share of population above 620 and implied approval rate gap. Source: author calculation and HMDA-McDash.

a *No Disparate Treatment* counterfactual, interpreted strictly; we will estimate the observed trade-offs between differences in denials and differences in default assuming that we loan if and only if borrower is above the threshold. The next will be a *No Disparate Impact* counterfactual, in which we require the bank loan to the top $k\%$ of borrowers in any group and consider the default rate trade-off. Finally we consider a *decoupled* regime - when we can set different strict threshold policies for different groups.

5.5.2 No Disparate Treatment

Now, we suppose we are in a strict No Disparate Treatment regime. That is, we must pick a *single* cutoff on credit scores, and loan to all applicants above this threshold and no applicants below it. The set of all thresholds represents our possible policy space under this regime; and so the outcomes we can achieve are constrained by the relationship between FICO score and default and distribution for each group. Figure 5.13 illustrates the these relationships. We construct the following estimates:

First, we simply estimate the approval rate for each group at a given FICO threshold. This is given theoretically by $\alpha_\tau(G) = \Pr[F \geq \tau | g = G]$, which is just the complement of the CDF; hence we estimate this quantity via $\hat{\alpha}_\tau(G) = 1 - \sum_{f \geq \tau} \widehat{\Pr}[F = f | G = g]$, the estimated CDF we obtained in 5.5.1. We also want to construct an overall weighted approval rate; we do this by simply constructing a

weighted average of $\hat{\alpha}(G)$, with the weights being the share of each group in HMDA.

Next, we estimate the probability that a given member of a given group at a given FICO score defaults. To do this, we construct:

$$\hat{\delta}(F, G, l) = \hat{\Theta}_G^{\text{HGBRT}}(F, l),$$

where $\hat{\Theta}_G^{\text{HGBRT}}$ is the monotonically constrained HGBRT we learned for each group in Section 5.4, and average over[25] $l \in \{80, ...99\}$. We define this average as $\hat{\delta}(F, G)$. Then we estimate the group default rate as:

$$\hat{\delta}_\tau(G) = \frac{\sum_{F:F\geq\tau} \widehat{\Pr}[f = F|g = G] \cdot \hat{\delta}(F, G)}{\sum_{F:F\geq\tau} \widehat{\Pr}[f = F|g = G]}$$

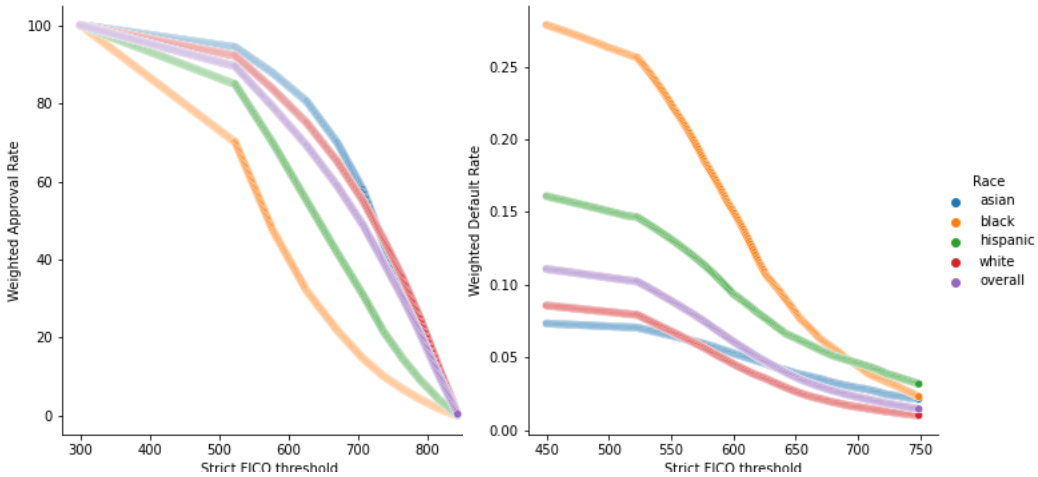
and similarly construct the overall default rate by weighting the race-specific rates by their prominence in HMDA. Importantly, our default rate estimates given a threshold τ only include borrowers who would receive a loan under this policy. Notice that the shape of this estimate will depend on the relationship of the threshold to the default distributions rather than the FICO itself.

The estimated results are plotted in Figure 5.13. As expected, Figure 5.13a is simply a reflected version of the CDF above, and shows that at *any* given fixed FICO threshold besides the trivial thresholds of the minimal and maximal FICO scores, we should expect Blacks and Hispanics to suffer significantly lower approval rates than Whites and Asians. Figure 5.13b displays estimated defaults for loans above each possible threshold. Qualitatively, the relationship is similar to that of approval but flipped: for the most part, every given FICO threshold will result in a higher default rate among Blacks and Hispanics than Whites and Asians. Finally, it is worth keeping in mind that the "Overall" figures are much closer to the line for Whites than the lines for Blacks or Hispanics. This is, of course, a trivial

[25] This is a simplification; a more sophisticated approach would be to, for instance, sample with replacement observations from the data so that we account for the conditional distribution of loan-to-value given the FICO score. But since, again, this work is intended to be somewhat stylized, we ignore this subtlety for now.

consequence of the fact that Whites are the majority. But what that means is that from a portfolio-wide (or system-wide) viewpoint, the difference between Blacks and Whites and Hispanics and Whites is not as costly in terms of total defaults as implied by looking at each group alone, and so lowering the lending threshold to expand access to credit may viable.

Now, using those two plots, we can plot the *Pareto* frontiers that show all the possibilities we can achieve in terms of both the overall denial and default rate, and the *gap* between groups in terms of denials and defaults. Figure 5.14 plots these frontiers by tracing out possible FICO thresholds and plotting their estimated $\hat{\alpha}$ and $\hat{\delta}$. In Figure 5.13a, we can see that the trade-off is not so surprising – the curve it self is convex, which suggests that there are diminishing returns in either direction. Figure 5.14b has an interesting shape - it is concave, and even loops back on itself (and so is not a function). That is because as we shift the threshold from low to high (starting from the top left and moving right and down along the curve) we can achieve similar differences in multiple ways. For instance, approving many Whites and a moderate number of Blacks can give the same disparity as approve a Moderate number of Whites and few Blacks; similarly, we can obtain a given disparity at various points along the absolute spectrum of default probability.

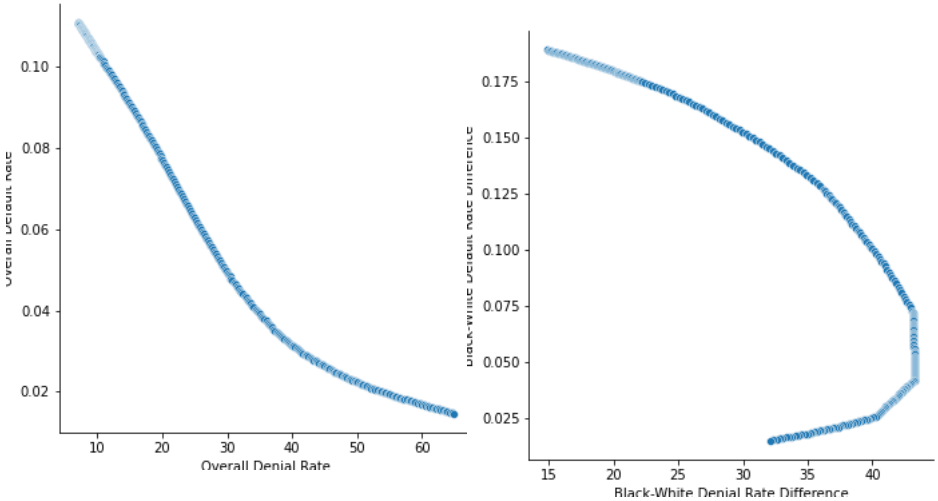


(a) Approval Source: Author calculation using HMDA-McDash and results of calculations using [107] and [74].

(b) Default rate Source: Author calculation using HMDA-McDash and results of calculations using [107] and [74].

Figure 5.13: Estimated approval rates and default rates in a (strict) No Disparate Treatment regime.

Finally, we note that both the differences and systemwide figures we estimate in Table 5.3 are on the *exterior* of these Pareto frontiers – that is, they strictly dominate some models on this frontier. (Again, these are only the Pareto frontiers for strictly No Disparate Treatment regimes.) This indicates that our real-world setting may be Pareto-improving over some alternatives.



(a) Tradeoff in default rate among approved borrowers vs. denial rate among applicants. Source: Author calculation using HMDA-McDash and results of calculations using [107] and [74].

(b) Trade-off in Black-White default gap vs Black-White denial gap. Source: Author calculation using HMDA-McDash and results of calculations using [107] and [74].

Figure 5.14: Trade-offs in denials vs default using strict FICO thresholds

Figure 5.15 illustrates the tradeoff between default rate and denial rate within each group and overall. Again, each point corresponds to a particular choice of τ and the estimated $1-\hat{\alpha}(G), \hat{\delta}(G)$, passing from very low denial rate and high default to very high denial rates and low defaults. The separation of these curves indicates that there are no points where the minority groups achieve the same level on both dimensions as the majority group – and this is what prevents us from achieving fairness on both metrics at the same time in a strictly disparate treatment regime.

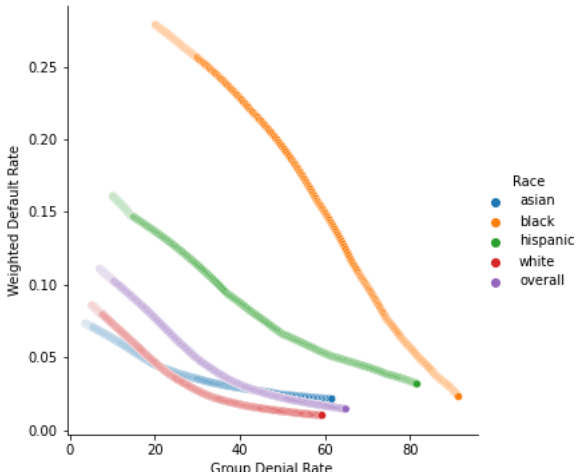


Figure 5.15: Denials vs Default Source: Author calculation using HMDA-McDash and results of calculations using [107] and [74].

5.5.3 No Disparate Impact

Now, we turn to No Disparate Impact regimes. Here, we can choose a fixed θ such that for each group, we take the top θ fraction of the distribution (which will correspond to different τ_Gs). In a slight abuse and overloading of notation, we can think of τ_G as a function mapping θ to some threshold t such that $\Pr[F \geq t|g = G] = \theta$. Because FICO scores are discrete, we can estimate τ_G in the following manner:

$$\hat{\tau}_G := \text{argmin}_{\underline{f}} f : \left[\sum_{F \geq \underline{f}} \widehat{\Pr}[f = F|g = G] \right] \leq \theta$$

We plot these estimates in Figure 5.16. Unsurprisingly, choosing θ close to 0 requires a very high credit score, while choosing it close to 100 requires very little in the way of a credit score at all. But here, too, we see a gap comparable to those previous – at any given θ, there is a large credit score gap between e.g. Blacks and Whites. So there is no place, besides again trivially at the top or bottom of the FICO scale, where we could set a cutoff that satisfies No Disparate Impact but also keeps a small difference in credit thresholds.

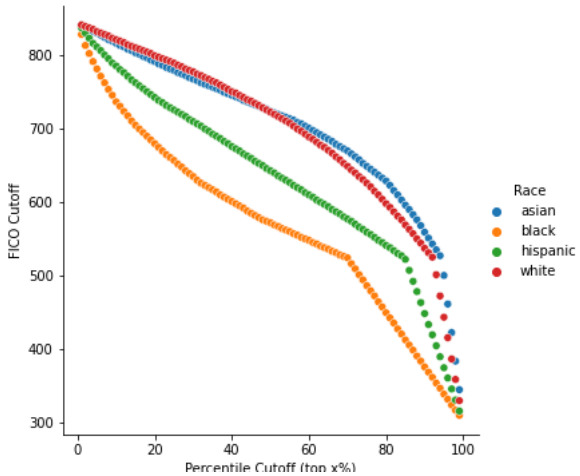
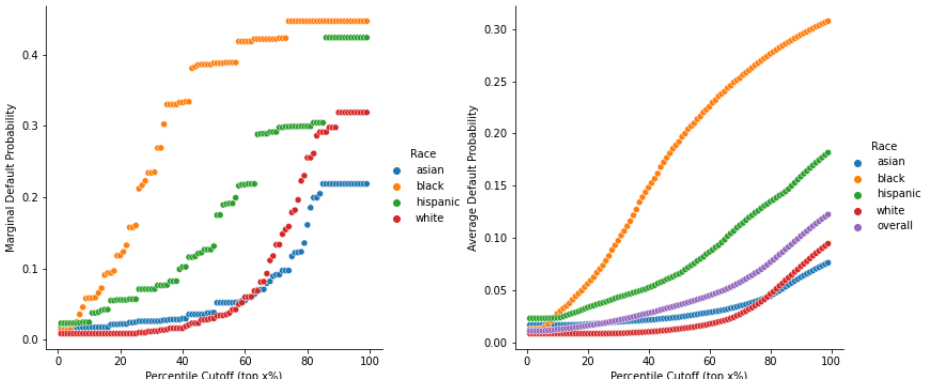


Figure 5.16: Denials vs Default. Source: Author calculation using HMDA-McDash and results of calculations using [107] and [74].

We see a similar result when we turn to defaults: both viewed at the margin, i.e. the default rate at $F = \tau_G$, as well as the overall default rates for $F \geq \tau_G$, there is a large gap beyond very high FICO scores. These figures are plotted in Figure 5.17. As before, the default curves of Blacks and Hispanics tend to be significantly higher at any given percentage point (viewed on both a marginal and averaged basis), but default rates are very similar among the top few percentage points of borrowers of each group. But as before, the fact that Whites are a larger fraction of the population overall means that the overall default rate tracks the White default curve more closely. Hence, higher fractions of borrower groups could be chosen while maintaining lower overall default rates than predicted by individual group default curves.



(a) Marginal DefaultSource: Author calculation using HMDA-McDash and results of calculations using [107] and[74].

(b) Average default. Source: Author calculation using HMDA-McDash and results of calculations using [107] and[74].

Figure 5.17: Default when taking top $x\%$ of each group.

Figure 5.18 displays the possible differences in White-Black FICO cutoff implied by a given percentile cutoff versus the Black-White difference in group default rate. The color of the points represent the percentile being accepted. The curve is best conceptualized as starting by accepting 0% and sweeping out the curve as the percentage increases, moving from near equal default rates with small FICO cutoff differences, to higher differences and default rate differences that grow together, and then eventually decreasing then increasing, then decreasing again in cutoff difference.

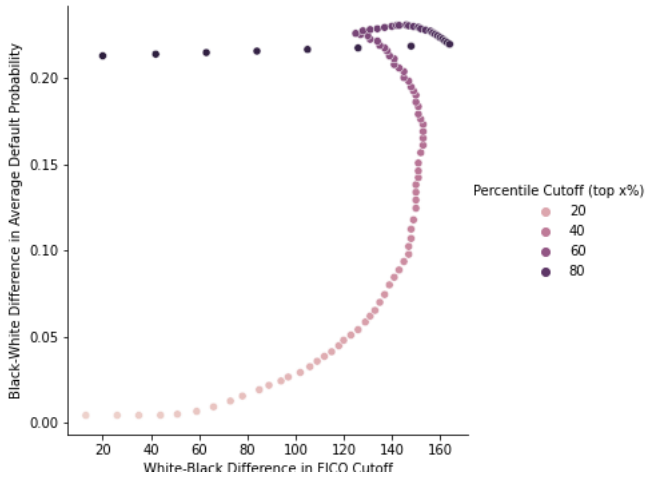


Figure 5.18: Denial gap-default gap trade-offs under No Disparate Impact. Source: Author calculation using HMDA-McDash and results of calculations using [107] and [74].

This figure highlights several features of the distribution. First, when extending credit to the very top applicants, there is very little difference between Black and White default rates. Instead, the gap is driven by borrowers nearer to the middle and beyond of their respective distributions. Second, any given gap in thresholds can usually be reached via multiple choices of the top-θ fraction. For instance, choosing to accept every applicant results in group FICO thresholds for each group of essentially the minimum FICO score, and similarly, choosing to accept almost no applicants in each group results in FICO cutoffs of essentially the maximum FICO score for each group. Third, as the cutoff of fraction increases from the top 1% to to the top 100%, both the default gap and the threshold difference may increase or decrease depending on where in the distribution one is.

5.5.4 Hybrid Regime

The trade-offs we have seen so far assume that banks are either satisfying a strict No Disparate Treatment or No Disparate Impact regime. If we instead allow them

to satisfy neither fully, we can obtain other possible choices in terms of default rates and denial rates. We consider here a *hybrid regime* in the following sense: within any group, there is a *strict* threshold policy on FICO score – similar to the strict thresholds of No Disparate Impact - - but we allow different thresholds for each group. (Note that if FICO distributions are fixed, we can equivalently think of these policies as taking a top fraction of each group θ_g, which differs by group.) Here, we consider the Black-White differences again, so we conceptualize a policy as a pair (τ_B, τ_W). In Figure 5.19, we evaluate the difference between Black-White default[26] rates and approval rates of pairs of policies in the cross product of the set $\{450, 455, ...750\} \times \{450, 455, ...750\}$.

The result is the feather-like pattern in Figure 5.19. In this picture, we have plotted each possible policy pair as a point, with its x-coordinate corresponding to the induced Black-White approval difference and its y-coordinate representing the Black-White default difference, while the color represents difference between the majority and minority cutoffs. The blue arrows indicate directions of increasing fairness along either dimension, and so point towards the origin, which is the point of total fairness as measured by these metrics. Hence to the left of 0, the arrow point right, but on the right of 0, the arrow points left, and so on. Finally, the black dotted lines approximately map the real-world Black-White default and approval differences, so their intersection is the point of decoupled policies that most closely matches our aggregate figures.

[26]We calculate weighted default rates as described in Section 5.4.2, and again average over uniform LTV between 80-100; repeating the exercise using the empirical LTV distribution gives very similar results.

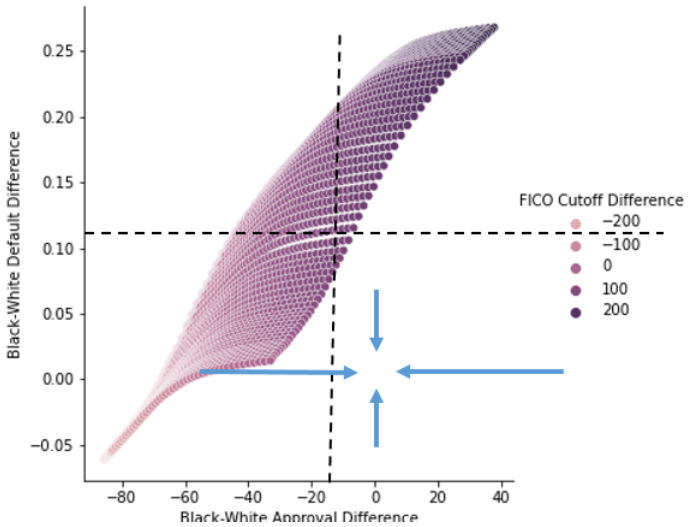


Figure 5.19: The Pareto"Feather" and associated frontier of possible decouple policies and their fairness measures. Black dotted lines intersect at our observed measures. Blue arrows point in the direction of more fairness. Source: Author calculation using HMDA-McDash and results of calculations using [107] and [74].

Interestingly, there are several policy pairs that appear to dominate our setting, but not very many (12 out of the 3600 in the space of policy pairs, and all of the same character. They keep the minority cutoff essentially fixed at 620 or up to 635, while *raising* the majority cutoff to 730-750. So such policies would gain fairness according to both metrics, but decrease the total loans made somewhat to the minority and significantly to the majority. On the other hand, there are plenty of policies that sacrifice on a single dimension but improve on another, and these may be preferable to where we are. Regardless, and perhaps surprisingly, our policy regime in practice appears to be near the Pareto frontier.

To the the extent we find its assumptions credible, Figure 5.19 appears to show us what is possible. For instance, there is no policy in the range we searched that leads to the $(0, 0)$ apparently-perfect fairness point. But in actuality, it shows us what is possible *today* – or even, since it is the result of learning on historical data, what was possible *yesterday*. As a society, we certainly should want to be on the

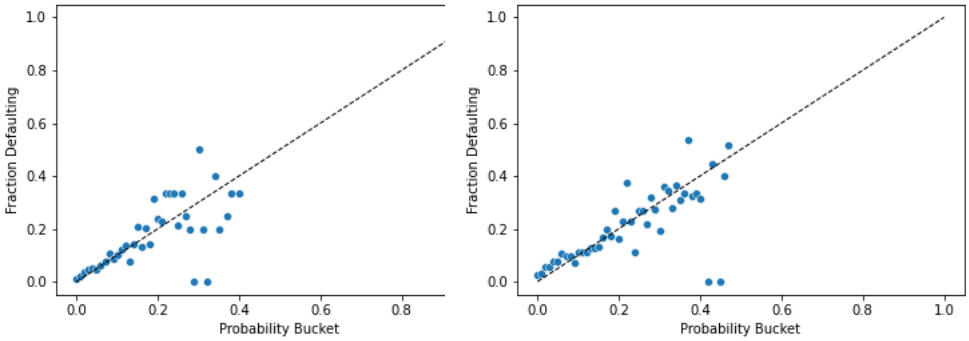
Pareto frontier, and where exactly we should be on that frontier is a question that deserves careful philosophical and moral debate. But ultimately, the goal should be to remedy historical injustices and push that Pareto frontier closer towards the point of perfect fairness – and then perhaps the curve of the frontier will not seem so cutting.

5.6 Conclusions and Future Work

In this Chapter, we have seen that the fairness measurements defined mathematically in the service of designing fairer algorithms can be used to quantify disparities in real-world settings, and we used these quantities to evaluate the current disparities in the mortgage market. But more interestingly, our counterfactual analysis lets us quantify and visualize the space of policy choices and their impacts, at least in the short term. Such an analysis can help make clear to policymakers and the public what choices we are implicitly making, and identify potential alternatives.

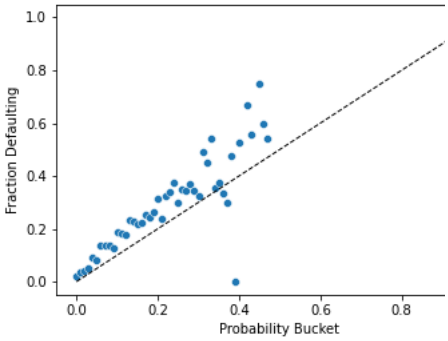
There are many limitations to our analysis, however. In particular, the assumptions we needed to make to derive some of our counterfactual results may be too coarse an approximation to reality, such as the case of the group-specific FICO distributions; in some other cases, such as the assumption that defaults among borrowers and defaults among the population are similar, we run a real risk of selection bias impacting our conclusions. To overcome these limitations, we can obtain exogeneity using experiments and natural experiments, focus on applicants with multiple applications, and also obtain richer soon-to-be-publicly-available data (e.g. HMDA 2018 and later data) or partner with credit ratings agencies or government institutions to get finer data. These improvements would strengthen the credibility of the results in this Chapter.

5.7 Appendix to Chapter 5

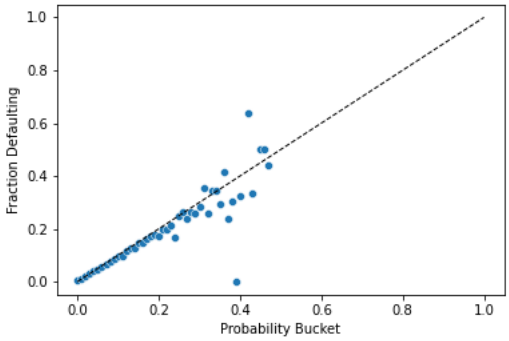


(a) Asian

(b) Black



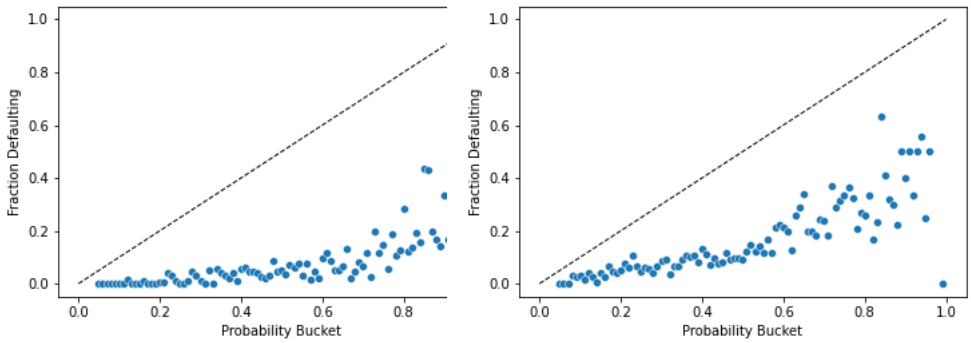
(c) Hispanic



(d) White

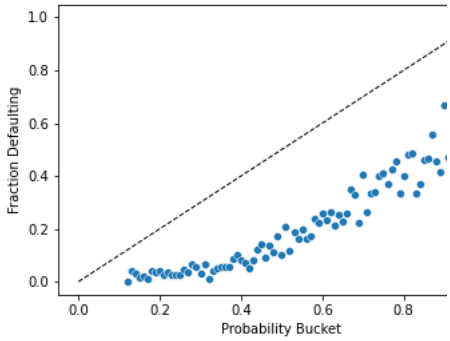
Figure 5.20: Default model calibration on test set (single model for all borrowers)

Source: Author calculation using HMDA-McDash.

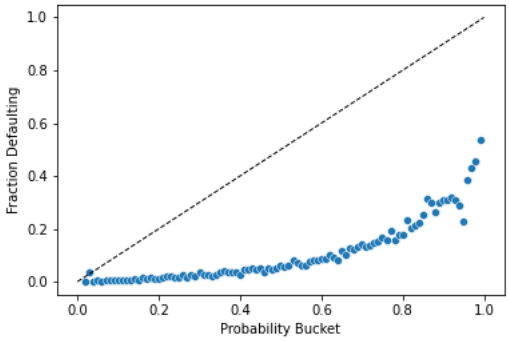


(a) Asian

(b) Black



(c) Hispanic



(d) White

Figure 5.21: Default model calibration in test set (separate logistic regression models) Source: Author calculation using HMDA-McDash.

Dep. Variable:	denied	**R-squared:**	0.034
Model:	OLS	**Adj. R-squared:**	0.034
Method:	Least Squares	**F-statistic:**	2452.
Date:	Fri, 20 Nov 2020	**Prob (F-statistic):**	0.00
Time:	02:15:02	**Log-Likelihood:**	-1.7487e+06
No. Observations:	4972008	**AIC:**	3.498e+06
Df Residuals:	4971936	**BIC:**	3.499e+06
Df Model:	71		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.4873	0.002	256.949	0.000	0.484	0.491
race[T.black]	0.1016	0.001	122.690	0.000	0.100	0.103
race[T.hisp]	0.0496	0.001	63.326	0.000	0.048	0.051
race[T.white]	-0.0262	0.001	-39.596	0.000	-0.028	-0.025
C(has_coap)[T.True]	-0.0401	0.000	-126.426	0.000	-0.041	-0.039
lti	-7.404e-05	8.9e-06	-8.315	0.000	-9.15e-05	-5.66e-05
np.power(lti, 2)	-2.024e-09	4.02e-10	-5.037	0.000	-2.81e-09	-1.24e-09
State FE included						
Year FE included						

Omnibus:	1673166.576	**Durbin-Watson:**	1.967
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4102318.803
Skew:	1.956	**Prob(JB):**	0.00
Kurtosis:	5.120	**Cond. No.**	4.84e+07

Table 5.7: OLS Regression Results Source: Author calculation using HMDA-McDash.

Dep. Variable:	denied	**No. Observations:**		4972008		
Model:	Logit	**Df Residuals:**		4971936		
Method:	MLE	**Df Model:**		71		
Date:	Fri, 20 Nov 2020	**Pseudo R-squ.:**		0.03700		
Time:	02:12:21	**Log-Likelihood:**		-1.9637e+06		
converged:	True	**LL-Null:**		-2.0392e+06		
	coef	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
Intercept	0.0174	0.012	1.422	0.155	-0.007	0.041
race[T.black]	0.6420	0.006	100.372	0.000	0.629	0.654
race[T.hisp]	0.3388	0.006	54.502	0.000	0.327	0.351
race[T.white]	-0.2320	0.006	-41.927	0.000	-0.243	-0.221
C(has_coap)[T.True]	-0.3529	0.003	-126.804	0.000	-0.358	-0.347
lti	-0.0005	7.13e-05	-7.241	0.000	-0.001	-0.000
np.power(lti, 2)	-1.497e-08	2.86e-09	-5.244	0.000	-2.06e-08	-9.38e-09
State FE included						
Year FE included						

Table 5.8: Logistic Regression Results Source: Author calculation using HMDA-McDash.

Chapter 6

CONCLUSION

This dissertation has explored several topics at the intersection of algorithms, markets, and society. These subjects are bona fide fields of inquiry in and of themselves, so we can only scratch the surface. Yet we believe the material covered here provides some useful insight. Of course, we believe the particular results achieved are valuable contributions to the problems for which they were derived:

- We have analyzed welfare and revenue in various auction formats relevant to a huge portion of the digital economy; disentangling allocation and pricing, we have gone beyond the standard model and mechanisms to analyze strategic considerations and implications for welfare and revenue.

- We have designed a new form of double auction which mathematically guarantees the preservation of privacy for participants' information; in theory, at least, this could greatly improve the efficiency of financial markets by allowing participants to forgo algorithmic strategies to preserve this privacy at a cost.

- We have applied economic and learning theory to show how even in the absence of technical factors, market forces can lead to unfairness in AI and machine learning-driven markets, which likely will be among the most societally impactful markets of the future; we have also shown that competition, a

canonical market-based panacea may not be such here, and that by contrast, regulation may be useful.

- We have adapted metrics from the Fair Machine Learning literature to one of the most societally important markets of the past and present – the U.S. mortgage market – to describe the empirical disparities in our world; we have also shown how to estimate the universe of choices and trade-offs we face in the short run.

To paraphrase a quote attributed to Kant: Practice without Theory is blind; Theory without Practice is lame. We have tried to ensure that our theoretical models do not stray too far from reality and that our practice be at least guided by theoretical justification. With this, we hope that our contributions will have real value and relevance. But beyond our specific contributions, the work here is of a particular perspective: the application of algorithmic tools and thinking, in conjunction with rigorous modeling of strategic agents and markets, and cognizance of, if not expertise in, social science and philosophical thinking, is a powerful combination. Given the increasing integration of algorithms into our society, we believe that this perspective will be necessary to understand and improve our rapidly changing world, and we are glad to promote it here.

BIBLIOGRAPHY

[1] Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.

[2] Abrams, Z., Ghosh, A., and Vee, E. Cost of conciseness in sponsored search auctions. In *Proc. of 3rd International Conference on Web and Internet Economics* (2007), pp. 326–334.

[3] Acharya, V., Richardson, M., van Nieuwerburgh, S., and White, L. *Guaranteed to Fail: Fannie Mae, Freddie Mac, and the Debacle of Mortgage Finance*. Princeton University Press, 2011.

[4] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).

[5] Agarwal, S., Benmelech, E., Bergman, N., and Seru, A. Did the community reinvestment act (cra) lead to risky lending? Tech. rep., National Bureau of Economic Research, 2012.

[6] Aridor, G., Liu, K., Slivkins, A., and Wu, Z. S. Competing bandits: The perils of exploration under competition. *arXiv preprint arXiv:1902.05590* (2019).

[7] Arrow, K. J. The theory of discrimination. In *In Discrimination in Labor Markets* (1973), Citeseer.

[8] Ausubel, L. M., and Milgrom, P. The lovely but lonely vickrey auction. In *Combinatorial Auctions, chapter 1* (2006), MIT Press.

[9] Banker, R. D., Khosla, I., and Sinha, K. K. Quality and competition. *Manage. Sci. 44*, 9 (Sept. 1998), 1179–1192.

[10] Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[11] Barocas, S., and Selbst, A. D. Big data's disparate impact. *Calif. L. Rev. 104* (2016), 671.

[12] Bartlett, R., Morse, A., Stanton, R., and Wallace, N. Consumer-lending discrimination in the fintech era. Tech. rep., National Bureau of Economic Research, 2019.

[13] Becker, G. The economics of discrimination. Tech. rep., University of Chicago Press, 1971.

[14] Ben-Porat, O., and Tennenholtz, M. Competing prediction algorithms. *arXiv preprint arXiv:1806.01703* (2018).

[15] Ben-Porat, O., and Tennenholtz, M. Regression equilibrium. *arXiv preprint arXiv:1905.02576* (2019).

[16] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.

[17] Berry, S. T. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics 25*, 2 (1994), 242–262.

[18] Bhutta, N. The community reinvestment act and mortgage lending to lower income borrowers and neighborhoods. *The Journal of Law and Economics 54*, 4 (2011), 953–983.

[19] Binns, R. Fairness in machine learning: Lessons from political philosophy. *arXiv preprint arXiv:1712.03586* (2017).

[20] Blodgett, S. L., and O'Connor, B. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061* (2017).

[21] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems* (2016), pp. 4349–4357.

[22] Brooks, T. *Rawls and Law*. Routledge, 2017.

[23] Bubb, R., and Kaufman, A. Securitization and moral hazard: Evidence from credit score cutoff rules. *Journal of Monetary Economics 63* (2014), 1–18.

[24] Budish, E., Cramton, P., and Shim, J. The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response. *The Quarterly Journal of Economics 130*, 4 (07 2015), 1547–1621.

[25] Buolamwini, J., and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency* (2018), pp. 77–91.

[26] Caragiannis, I., Kaklamanis, C., Kanellopoulos, P., Kyropoulou, M., Lucier, B., Leme, R. P., and Tardos, É. Bounding the inefficiency of outcomes in generalized second price auctions. *Journal of Economic Theory 156* (2015), 343–388.

[27] Cavallo, R., Sviridenko, M., and Wilkens, C. A. Matching auctions for search and native ads. In *Proceedings of the 2018 ACM Conference on Economics and Computation* (2018), pp. 663–680.

[28] Cavallo, R., and Wilkens, C. A. Gsp with general independent click-through-rates. In *Proc. of 10th International Conference on Web and Internet Economics* (2014), T.-Y. Liu, Q. Qi, and Y. Ye, Eds., pp. 400–416.

[29] Challet, D. Strategic behaviour and indicative price diffusion in paris stock exchange auctions. In *New Perspectives and Challenges in Econophysics and Sociophysics*. Springer, 2019, pp. 3–12.

[30] Challet, D., and Gourianov, N. Dynamical regularities of us equities opening and closing auctions. *Market Microstructure and Liquidity 4*, 1 (2018).

[31] Chawla, S., and Hartline, J. D. Auctions with unique equilibria. In *Proceedings of the fourteenth ACM conference on Electronic commerce* (2013), pp. 181–196.

[32] Chen, I., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems* (2018), pp. 3539–3550.

[33] Chen, Z., Ni, T., Zhong, H., Zhang, S., and Cui, J. Differentially private double spectrum auction with approximate social welfare maximization. *arXiv preprint arXiv:1810.07873* (2018).

[34] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5*, 2 (2017), 153–163.

[35] Clarke, E. H. Multipart pricing of public goods. *Public choice 11*, 1 (1971), 17–33.

[36] Colini-Baldeschi, R., Leonardi, S., Schrijvers, O., and Sodomka, E. Envy, regret, and social welfare loss. In *Proceedings of The Web Conference 2020* (2020), pp. 2913–2919.

[37] Colini-Baldeschi, R., Mestre, J., Schrijvers, O., and Wilkens, C. A. The ad types problem. In *Web and Internet Economics: 16th International Conference, WINE 2020, Ljublana, Slovenia, December 17–20, 2017, Proceedings* (2020), Springer.

[38] Cowgill, B., and Tucker, C. E. Economics, fairness and algorithmic bias. *preparation for: Journal of Economic Perspectives* (2019).

[39] Cummings, R., Kearns, M., Roth, A., and Wu, Z. S. Privacy and truthful equilibrium selection for aggregative games. In *International Conference on Web and Internet Economics* (2015), Springer, pp. 286–299.

[40] Diana, E., Elzayn, H., Kearns, M., Roth, A., Sharifi-Malvajerdi, S., and Ziani, J. Differentially private call auctions and market impact. In *Proceedings of the 21st ACM Conference on Economics and Computation* (New York, NY, USA, 2020), EC '20, Association for Computing Machinery, p. 541–583.

[41] Dobbie, W., Liberman, A., Paravisini, D., and Pathania, V. Measuring bias in consumer lending. Tech. rep., National Bureau of Economic Research, 2018.

[42] Dong, J., Elzayn, H., Jabbari, S., Kearns, M., and Schutzman, Z. Equilibrium characterization for data acquisition games.

[43] Dong, J., Roth, A., Schutzman, Z., Waggoner, B., and Wu, Z. S. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference on Economics and Computation* (2018), pp. 55–70.

[44] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (2012), pp. 214–226.

[45] Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography* (Berlin, Heidelberg, 2006), S. Halevi and T. Rabin, Eds., Springer Berlin Heidelberg, pp. 265–284.

[46] Dwork, C., and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science 9*, 3–4 (2014), 211–407.

[47] Dwork, C., Rothblum, G. N., and Vadhan, S. Boosting and differential privacy. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science* (Washington, DC, USA, 2010), FOCS '10, IEEE Computer Society, pp. 51–60.

[48] Edelman, B., Ostrovsky, M., and Schwarz, M. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review 97*, 1 (2007), 242–259.

[49] Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., and Pera, M. S. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on Fairness, Accountability and Transparency* (2018), pp. 172–186.

[50] Elzayn, H., and Fish, B. The effects of competition and regulation on error inequality in data-driven markets. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2020), FAT* '20, Association for Computing Machinery, p. 669–679.

[51] Elzayn, H., Jabbari, S., Jung, C., Kearns, M., Neel, S., Roth, A., and Schutzman, Z. Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), ACM, pp. 170–179.

[52] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847* (2017).

[53] Even-Dar, E., Kakade, S. M., Kearns, M., and Mansour, Y. (in)stability properties of limit order dynamics. In *Proceedings of the 7th ACM Conference on Electronic Commerce* (New York, NY, USA, 2006), EC '06, Association for Computing Machinery, p. 120–129.

[54] Feldman, M., Lucier, B., and Nisan, N. Correlated- and coarse- equilibria of single-item auctions. *CoRR abs/1601.07702* (2016).

[55] Feng, Z., Guruganesh, G., Liaw, C., Mehta, A., and Sethi, A. Convergence analysis of no-regret bidding algorithms in repeated auctions, 2020.

[56] Fish, B., Bashardoust, A., Boyd, D., Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. Gaps in information access in social networks. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019* (2019), pp. 480–490.

[57] Fishback, P., LaVoice, J., Shertzer, A., and Walsh, R. Race, risk, and the emergence of federal redlining. Tech. rep., National Bureau of Economic Research, 202-.

[58] Flores, A. W., Bechtel, K., and Lowenkamp, C. T. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's

software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation 80* (2016), 38.

[59] Fudenberg, D., Tirole, J., TIROLE, J., and Press, M. *Game Theory*. Mit Press. MIT Press, 1991.

[60] Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets (November 6, 2018)* (2018).

[61] Gatheral, J. No-dynamic-arbitrage and market impact. *Quantitative finance 10*, 7 (2010), 749–759.

[62] Gatheral, J. Three models of market impact. In *Market Microstructure and High Frequency Data* (2010).

[63] Gomes, R., and Sweeney, K. Bayes–nash equilibria of the generalized second-price auction. *Games and economic behavior 86* (2014), 421–437.

[64] Greenlaw, P. S., and Jensen, S. S. Race-norming and the civil rights act of 1991. *Public personnel management 25*, 1 (1996), 13–24.

[65] Groves, T. Incentives in teams. *Econometrica: Journal of the Econometric Society* (1973), 617–631.

[66] Guryanov, A. Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees. In *International Conference on Analysis of Images, Social Networks and Texts* (2019), Springer, pp. 39–50.

[67] Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (2016), pp. 3315–3323.

[68] Hartline, J., Syrgkanis, V., and Tardos, E. No-regret learning in bayesian games. In *Advances in Neural Information Processing Systems* (2015), pp. 3061–3069.

[69] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[70] Hayashi, F. *Econometrics*. Princeton University Press, 2011.

[71] Heidari, H., Ferrari, C., Gummadi, K., and Krause, A. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems* (2018), pp. 1265–1276.

[72] Hsu, J., Huang, Z., Roth, A., Roughgarden, T., and Wu, Z. S. Private matchings and allocations. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing* (New York, NY, USA, 2014), STOC '14, Association for Computing Machinery, p. 21–30.

[73] Hsu, J., Huang, Z., Roth, A., and Wu, Z. S. Jointly private convex programming. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms* (2016), SIAM, pp. 580–599.

[74] Huynh, F. Fico® score distribution remains mixed, 2013.

[75] Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in reinforcement learning. In *International Conference on Machine Learning* (2017), PMLR, pp. 1617–1626.

[76] Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems* (2016), pp. 325–333.

[77] Judd, K. L. Cournot versus bertrand: A dynamic resolution.

[78] Kallus, N., and Zhou, A. Residual unfairness in fair machine learning from prejudiced data. *arXiv preprint arXiv:1806.02887* (2018).

[79] Kannan, S., Morgenstern, J., Roth, A., and Wu, Z. S. Approximately stable, school optimal, and student-truthful many-to-one matchings (via differential privacy). In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms* (2014), SIAM, pp. 1890–1903.

[80] Kaplan, T. R., and Zamir, S. Asymmetric first-price auctions with uniform distributions: analytic solutions to the general case. *Economic Theory 50*, 2 (2012), 269–302.

[81] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (2017), pp. 3146–3154.

[82] Kearns, M. Fair algorithms for machine learning. In *Proceedings of the 2017 ACM Conference on Economics and Computation* (New York, NY, USA, 2017), EC '17, ACM, pp. 1–1.

[83] Kearns, M., Pai, M. M., Roth, A., and Ullman, J. Mechanism design in large games: Incentives and privacy. *The American Economic Review 104*, 5 (2014), 431–435.

[84] Kearns, M., and Roth, A. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, Incorporated, 2019.

[85] Kearns, M. J., and Vazirani, U. V. *An Introduction to Computational Learning Theory*. Mit Press. MIT Press, 1994.

[86] Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[87] Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly 2*, 1-2 (1955), 83–97.

[88] Leme, R. P., and Tardos, E. Pure and bayes-nash price of anarchy for generalized second price auction. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science* (2010), IEEE, pp. 735–744.

[89] Lewis, M. *Flash Boys: A Wall Street Revolt*. A Wall Street Revolt. W. W. Norton, 2014.

[90] Lucier, B., and Paes Leme, R. GS0p auctions with correlated types. In *Proceedings of the 12th ACM conference on Electronic commerce* (2011), pp. 71–80.

[91] Lum, K., and Isaac, W. To predict and serve? *Significance 13*, 5 (2016), 14–19.

[92] Mac, F.

[93] Mansour, Y., Slivkins, A., and Wu, Z. S. Competing bandits: Learning under competition. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)* (2018), Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[94] McSherry, F., and Talwar, K. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science* (Washington, DC, USA, 2007), FOCS '07, IEEE Computer Society, pp. 94–103.

[95] Mehrotra, R., Anderson, A., Diaz, F., Sharma, A., Wallach, H., and Yilmaz, E. Auditing search engines for differential satisfaction across demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Republic and Canton of Geneva, Switzerland, 2017),

WWW '17 Companion, International World Wide Web Conferences Steering Committee, pp. 626–633.

[96] Munkres, J. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics 5*, 1 (1957), 32–38.

[97] Narayanan, A. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA* (2018).

[98] NASDAQ. Nasdaq opening and closing crosses.

[99] NYSE. Nyse opening and closing auctions fact sheet.

[100] Obermeyer, Z., and Mullainathan, S. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), ACM, pp. 89–89.

[101] Parsons, S., Marcinkiewicz, M., Niu, J., and Phelps, S. Everything you wanted to know about double auctions, but were afraid to (bid or) ask.

[102] Parsons, S., Rodriguez-Aguilar, J. A., and Klein, M. Auctions and bidding: A guide for computer scientists. *ACM Computing Surveys (CSUR) 43*, 2 (2011), 1–59.

[103] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[104] Phelps, E. S. The statistical theory of racism and sexism. *The american economic review 62*, 4 (1972), 659–661.

[105] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems* (2017), pp. 5680–5689.

[106] Rawls, J. *A Theory of Justice*. Oxford University Press, 1999.

[107] Reserve, U. F. Report to the congress on credit scoring and its effects on the availability and affordability of credit. *Board of Governors of the Federal Reserve System* (2007).

[108] Rogers, R., Roth, A., Ullman, J., and Wu, Z. S. Inducing approximately optimal flow using truthful mediators. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation* (2015), pp. 471–488.

[109] Rogers, R. M., and Roth, A. Asymptotically truthful equilibrium selection in large congestion games. In *Proceedings of the Fifteenth ACM conference on Economics and Computation* (2014), pp. 771–782.

[110] Rothstein, R. *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright, 2017.

[111] Roughgarden, T. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM) 62*, 5 (2015), 1–42.

[112] Roughgarden, T., Syrgkanis, V., and Tardos, E. The price of anarchy in auctions. *Journal of Artificial Intelligence Research 59* (2017), 59–101.

[113] Shalev-Shwartz, S., and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[114] Shapiro, C. Theories of oligopoly behavior. *Handbook of industrial organization 1* (1989), 329–414.

[115] Simoiu, C., Corbett-Davies, S., Goel, S., et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics 11*, 3 (2017), 1193–1216.

[116] Sweeney, L. Discrimination in online ad delivery. *Queue 11*, 3 (Mar. 2013), 10:10–10:29.

[117] Syrgkanis, V., and Tardos, E. Composable and efficient mechanisms. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing* (2013), pp. 211–220.

[118] Taylor, K. *Race for Profit: How Banks and the Real Estate Industry Undermined Black Homeownership*. Justice, Power, and Politics. University of North Carolina Press, 2019.

[119] Tirole, J. *The theory of industrial organization*. MIT press, 1988.

[120] Tullock, G. Efficient rent seeking. In *Efficient Rent-Seeking*. Springer, 2001, pp. 3–16.

[121] Varian, H. R. Position auctions. *international Journal of industrial Organization 25*, 6 (2007), 1163–1178.

[122] Verma, S., and Rubin, J. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018* (2018), pp. 1–7.

[123] Vickrey, W. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of finance 16*, 1 (1961), 8–37.

[124] Wah, E., and Wellman, M. Latency arbitrage, market fragmentation, and efficiency: A two-market model. In *Proceedings of the Thirteenth ACM conference on Economics and Computation* (2013).

[125] Wilson, B., Hoffman, J., and Morgenstern, J. Predictive inequity in object detection. *CoRR abs/1902.11097* (2019).

[126] Wilson, B., Hoffman, J., and Morgenstern, J. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097* (2019).

[127] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning* (2013), pp. 325–333.